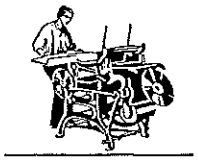
F.W. Lancaster

Indexação e resumos: teoria e prática

Segunda edição

Tradução de Antonio Agenor Briquet de Lemos



BRIQUET DE LEMOS
LIVROS

Título original: *Indexing and abstracting in theory and practice*

Primeira edição original: 1991
Segunda edição original: 1998
Terceira edição original: 2003

Primeira edição brasileira: 1993
Segunda edição brasileira: 2004 (baseada na terceira edição original de 2003)

Revisão: Maria Lucia Vilar de Lemos

Capa: Formatos Design e Informática Ltda.

Dados internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Lancaster, F.W., 1933–
Indexação e resumos : teoria e prática / F.W. Lancaster ; tradução de Antonio Agenor Briquet de Lemos. — 2. ed. rev. atual. — Brasília, DF : Briquet de Lemos / Livros, 2004.

Título original: Indexing and abstracting in theory and practice.
Bibliografia.

ISBN 85-85637-24-2

1. Indexação. 2. Resumos – Redação I. Título.

04-5201 CDD 025.3

Índices para catálogo sistemático
1. Indexação : Ciência da informação 025.3

2004

Briquet de Lemos / Livros
SRTS - Quadra 701 - Bloco K - Sala 831
Edifício Embassy Tower
Brasília, DF 70340-000
Telefones (61) 322 9806 / 313 6923
Fax (61) 323 1725
www.briquetdelemos.com.br
editora@briquetdelemos.com.br

Sumário

Para Shane, Aaron,
Rachael, Maddie, Alex,
Joshua, Evan e Emma,
bem como
Lakshmi e Rajeshwari

Prefácio

A primeira edição desta obra, que recebeu o prêmio de melhor livro do ano sobre ciência da informação, outorgado pela American Society for Information Science, foi publicada em 1991; a segunda foi lançada em 1998. Ambas foram bem-recebidas pelos críticos, e o livro tem sido amplamente utilizado como texto didático na América do Norte, no Reino Unido e em outros países.

Entre 1991 e 1998 este campo passou por mudanças notáveis, o que suscitou a necessidade de novos capítulos, principalmente sobre a internet e a indexação e elaboração de resumos para bases de dados de imagens e sons. As mudanças verificadas a partir de 1998 foram menos marcantes. No entanto, ocorreram avanços que definiam a necessidade de uma terceira edição.

Todo o texto foi atualizado, embora os capítulos iniciais, que tratam mais de princípios básicos, permaneçam bem similares aos da segunda edição. Em compensação, alguns dos capítulos finais foram substancial ou completamente reescritos. Refiro-me aos capítulos 13–17 que tratam, respectivamente, de bases de dados de imagens e sons, buscas em textos, indexação automática e atividades afins, indexação e a internet, e o futuro da indexação e da redação de resumos.

Não alterei muitas das figuras porque acho que as que foram utilizadas na segunda edição ainda continuam totalmente válidas para ilustrar os aspectos que desejo mostrar. Isso é ainda mais verdadeiro no que tange ao capítulo 10, sobre serviços impressos de indexação e resumos. Embora pudesse ter atualizado as páginas apresentadas como amostras, pareceu-me bastante desnecessário fazê-lo.

Embora a indexação e redação de resumos fossem antigamente tidas como processos que somente interessavam a bibliotecas e a algumas editoras, sua relevância e utilidade são reconhecidas hoje em dia de modo muito mais amplo, pois, obviamente, encontram aplicação em todos os tipos de recursos de informação em formato digital. Assim, esta edição, embora continue sendo destinada fundamentalmente ao uso como texto didático em escolas de biblioteconomia e ciência da informação (e programas afins), ainda se reveste de interesse para um público muito maior: produtores de bases de dados de todos os tipos, bem como aquelas pessoas interessadas em outras áreas, como o projeto de intranets, desenvolvimento de portais, sistemas de gerenciamento da informação, e gestão do conhecimento em geral.

Acho que devo dizer algo acerca das fontes citadas. O autor de uma recensão da primeira edição criticou-me por continuar citando fontes 'antigas'. Apesar de ter feito um esforço para atualizar por completo as fontes citadas (até o começo de 2003), não tenho por que me desculpar por continuar citando material antigo e até muito antigo. Para mim é inconcebível que um livro sobre este assunto deixe de citar (por exemplo) Cutter (1876) e Ranganathan (década de

1930). Ademais, muitas pessoas que hoje escrevem sobre esses temas parecem não ter interesse nem conhecer as primeiras contribuições feitas a este campo. Acredito que seja importante, principalmente para os estudantes, compreender como este campo se desenvolveu e reconhecer que muitas das idéias atualmente apresentadas como novas podem ser encontradas, de fato, na literatura de trinta ou mais anos passados, em forma um tanto similar.

Do mesmo que nas edições anteriores, esta não procura lidar com os índices de livros isolados, que aparecem no final dos livros impressos. Trata-se de assunto bem estudado em outras obras escritas por pessoas com muito mais experiência do que eu nessa área específica.

Esta edição deve ainda ser vista como um texto de natureza introdutória. Embora creia que os capítulos 1–12 sejam bastante abrangentes, já sobre os temas focalizados nos capítulos 13–15 foram escritos livros completos, de modo que esses capítulos, em particular, devem ser lidos como introduções a esses temas.

F.W. LANCASTER
Urbana, Illinois (EUA)
Março de 2003

AGRADECIMENTOS

Encontra-se consignada nas legendas das figuras a permissão para utilização de várias figuras de diferentes fontes. Além disso, quero agradecer a: Elsevier Science pela permissão para citar alguns trechos extensos de textos publicados em *Information Processing and Management*; OCLC Inc. pela permissão para reproduzir longas passagens de um artigo de O'Neill et al. (2001); John Wiley and Sons pela permissão para citar vários trechos extensos de material publicado no *Journal of the American Society for Information Science and Technology* (e seus antecessores); Information Today Inc. (<www.infotoday.com> pela permissão para reproduzir extensas citações de Hock (2001), de *EContent* e de *Online*; IBM pela permissão para reproduzir uma longa citação do *IBM Systems Journal*; Thomas Craven pela permissão para reproduzir citações de vários de seus artigos; Getty Research Institute por extensas citações de Layne (2002); IOS Press pela permissão de reproduzir uma extensa citação de Nielsen (1997); e ACM Publications pela permissão de fazer citação de Wactlar et al. (2002).

Os termos e definições extraídos da ISO 5963:1985 são reproduzidos com a permissão da International Organization for Standardization (ISO). Esta norma pode ser obtida junto a qualquer membro da ISO e no sítio na Rede da secretaria central da ISO no seguinte endereço: <www.iso.org>. O detentor do direito autoral é a ISO.

Por fim, quero agradecer a várias pessoas por sua ajuda nesta edição: Bella Weinberg por ter me chamado a atenção para algumas fontes que, de outra forma, me teriam passado despercebidas; Bryan Heidorn por ter lido um primeiro rascunho do capítulo 13; Susanne Humphrey e Lou Knecht por atualizarem as informações de que dispunha acerca da National Library of Medicine; June Silvester, do Center for AeroSpace Information; Chandra Prabha pelas informações do OCLC; o pessoal da Library and Information Science Library da University of Illinois (e especialmente Sandy Wolf), por sua paciente ajuda na localização de material para mim, e Kathy Painter pelo seu trabalho, tradicionalmente excelente, de colocar a revisão do texto em formato eletrônico.

F.W. LANCASTER
Urbana, Illinois
Abril de 2003

UMA NOTA SOBRE TERMINOLOGIA
(e a redescoberta da roda)

Tenho trabalhado em bibliotecas ou em torno delas há muitos anos. Durante grande parte desse tempo estive envolvido, de uma ou outra forma, com a análise de assuntos. Em 1957, comecei a trabalhar redigindo resumos, que abrangiam uma ampla gama de material científico e tecnológico, para um boletim de resumos para a indústria, tarefa que exigia também um nível minucioso de indexação temática dos itens resumidos. Em 1958, assumi o trabalho de editor desse boletim. Anteriormente tivera experiência com a classificação de livros numa biblioteca pública, além de redigir anotações, sobre características locais, a serem incluídas nas fichas catalográficas (na década de 1950 a catalogação cooperativa ou centralizada ainda não era a norma). Por volta de 1961 estava envolvido no campo da 'recuperação da informação', e publiquei meu primeiro artigo em 1963 e o primeiro livro em 1968.

Em outras palavras, tem sido muito longa minha participação nas áreas de análise temática/recuperação da informação, presenciei inúmeras mudanças e conheci muitos dos principais atores deste palco em particular.

Até o final da década de 1940 e começo da década de 1950, o campo que hoje lembramos como 'recuperação da informação' era domínio quase exclusivo da profissão de bibliotecário. A realização de duas importantes conferências internacionais, além do reconhecimento de que os computadores poderiam aportar uma contribuição importante ao problema da recuperação da informação, tornaram o campo mais atraente e para ele acorreram pesquisadores de muitas outras áreas.

Ao longo de um período de mais de 50 anos, as contribuições à bibliografia sobre recuperação da informação tiveram origem em praticamente todos os campos acadêmicos, inclusive matemática, ciência da computação, psicologia, estatística, direito e medicina (informática médica).

Embora rostos novos e novos enfoques sejam sempre bem-vindos, é lamentável que muitos dos que hoje trabalham neste campo não tenham nenhuma formação prévia e, por isso, nenhum alicerce sólido sobre o qual construir. O maior problema é causado pelo fato de que muitos dos que atualmente trabalham com recuperação da informação parecem completamente ignorantes do fato de que outros processos diferentes dos totalmente automáticos foram aplicados, com algum sucesso, à recuperação da informação durante mais de 100 anos, e que de fato existe uma bibliografia sobre recuperação da informação além daquela da comunidade de informática. Exemplo gritante encontra-se em Agosti et al. (1995), que definem as 'etapas da indexação' como "extração de termos [*term extraction*], remoção de termos proibidos [*stop-term removal*], fusão [*conflation*] e ponderação [*weighting*]".

Muitas idéias surgidas hoje possuem claros antecedentes na literatura de 30 ou 40 anos atrás, mas esses trabalhos pioneiros são completamente desconhecidos para os pesquisadores atuais. Um caso pertinente é a pesquisa sobre mapas visuais ou 'navegadores' [*browsers*] para facilitar a navegação em sistemas de hipermídia (por exemplo, Fowler et al., 1996; Zizi, 1996) que é basicamente uma redescoberta dos 'mapas semânticos' [*semantic road maps*] de Doyle (1961).

O campo da recuperação de imagens parece ser o pior de todos em matéria de reinventar a roda. Por exemplo, um artigo de Schreiber et al. (2001) descreve um esquema para indexação de fotografias (denominam-no 'anotação fotográfica baseada na ontologia' [*ontology-based photo annotation*], que se baseia essencialmente num conjunto bastante simples de facetas. Parece que acreditam que a análise de facetas surgiu com eles ou, pelo menos, com outros que trabalham na mesma área. Ironicamente, o trabalho deles foi publicado num periódico dedicado a 'sistemas inteligentes'.

Os cientistas da computação que escrevem sobre recuperação da informação parecem reconhecer e citar somente outros cientistas da computação que escrevem sobre recuperação da informação. Exemplo óbvio é o reconhecimento e a citação quase unânimes de Salton como a autoridade em medidas de revocação e precisão na avaliação de atividades de recuperação da informação. Gerard Salton, por mais importante que tenha sido no campo da recuperação da informação, com a maior certeza não foi o introdutor dessas medidas, que, de fato, remontam à década de 1950.

Esse fenômeno de redescobrimento foi salientado por Holmes (2001), ele próprio um cientista da computação, que nos faz lembrar a advertência feita por George Santayana para quem aqueles que não podem recordar o passado estão condenados a repeti-lo. Holmes, partindo disso, acrescenta:

> [...] o que pensamos que sejam inovações muitas vezes são meras repetições [...] nossa profissão pode desenvolver-se de modo mais rápido e melhor por meio de inovações cumulativas, construindo sobre os alicerces de seu passado ao invés de ignorá-lo (p. 144).

Ele afirma que, em particular, as obras de Vannevar Bush e Hans Peter Luhn, que datam de 40 ou 60 anos, contêm idéias que desde então são reinventadas.

Minha pior experiência com esse problema específico ocorreu há vários anos, quando deparei com um artigo escrito por um cientista europeu, essencialmente um matemático, acerca de assunto sobre o qual eu publicara anteriormente. Quando escrevi para mostrar que ele deixara de citar meu trabalho anterior, e diversos outros de autoria de outros pesquisadores, ele contestou, folgadamente, para dizer que nunca pesquisava na literatura, a não ser que estivesse escrevendo um artigo de revisão! Que espécie de não-ciência egoísta é essa?

Outro resultado da multiplicidade de profissões que agora contribuem para a literatura de análise temática/recuperação da informação está na substituição, sem necessidade, da terminologia, apropriada e reconhecida, da profissão bi-

bliotecária. Exemplo óbvio é 'metadados'. O Oxford English Dictionary (em linha) registra 1968 como o ano do aparecimento dessa palavra. Na época foi usada para designar dados que descreviam conjuntos de dados (numéricos ou estatísticos). Desde então tornou-se praticamente um substituto para 'descrição bibliográfica', denominação esta perfeitamente razoável, com a qual convivíamos há muitos e muitos anos e que é aceita em normas internacionais. Alguém, é claro, poderia argumentar que 'bibliográfico' aplica-se apenas a livros. Sua extensão, porém, a outras formas documentárias (como em 'base de dados bibliográficos' e 'referência bibliográfica') convive conosco há muito tempo.

Alguns autores, com certeza, chamaram atenção para o mesmo problema. Milstead e Feldman (1999), por exemplo, argumentam convincentemente:

> Quer o chamemos de catalogação, indexação ou metadados, o conceito é familiar aos profissionais da informação. Agora, o mundo eletrônico por fim o descobriu. Faz alguns anos, somente uns poucos filósofos haviam ouvido falar em 'metadados'. Hoje em dia, é difícil encontrar uma publicação sobre recursos eletrônicos que ignore essa palavra. [...] Como o personagem que passou toda a vida escrevendo prosa sem saber que o fazia,* os bibliotecários e indexadores vêm há séculos produzindo e normalizando metadados. Ignorando este legado, uma imensa variedade de outros atores ingressaram recentemente nesse campo, e muitos deles não têm qualquer idéia de que alguém mais antes deles já tenha 'estado ali, feito aquilo'. Sistemas diferentes estão sendo desenvolvidos para tipos diferentes — e às vezes os mesmos — de informação, disso resultando uma atmosfera caótica de normas conflitantes (p. 25).

Não obstante, parecem dispostas a aceitar a nova terminologia.

Pessoas de nosso próprio campo, que certamente deveriam saber mais (e ser mais responsáveis), colaboram com essa situação. Por exemplo, Greenberg (2003) nos diz que a geração de metadados por seres humanos ocorre quando uma pessoa, como um criador profissional de metadados ou um fornecedor de conteúdo, produz metadados. Para ela 'criador profissional de metadados' é o 'catalogador' ou 'indexador', conforme admite depois em seu artigo (embora ela também inclua '*web master*' nesta categoria). Fiquei profundamente chocado (e de modo algum satisfeito) ao saber que gastei vários anos de minha vida como criador profissional de metadados, se bem que inocente disso.

Muitos que escrevem sobre recuperação de imagens usam o termo 'anotação' para designar a atribuição de rótulos de texto, como palavras-chave, que identificam o que a imagem representa, o que, evidentemente, é 'indexação'. Isso é duas vezes lamentável porque 'anotação' [*annotation*], há muitos anos, é empregada para designar o que é, fundamentalmente, um resumo muito sucinto (que aparecia antigamente em fichas de catálogos). Liu e Li (2002) mencionam termos de indexação atribuídos a videoclipes como 'etiquetas de anotação' [*annotation tags*]. Parece que elas constituem uma 'descrição semântica' [*semantic description*] e são obtidas por meio de 'extração semântica' [*semantic extraction*] que, provavelmente, significa identificação do assunto tratado.

Parte dessa confusão terminológica se deve a desleixo no trabalho editorial. Faz pouco deparei com um artigo em que a palavra '*indexation*', que estava até no título, era usada como sinônimo de '*indexing*'. O vocábulo '*indexation*' realmente existe na língua inglesa, mas empregado apenas em contexto econômico (por exemplo, em relação a certas variáveis, como aumento ou redução de salários e juros às mesmas taxas do índice de custo de vida); quase com certeza não é sinônimo de '*indexing*'. Os autores, neste caso, têm uma desculpa porque são franceses ('*indexation*' é o equivalente francês de '*indexing*'), mas não há desculpa para os editores de um periódico em língua inglesa se permitirem tal incorreção. Aguardo agora que a palavra '*indexation*' venha a substituir '*indexing*' na literatura de ciência da computação.

Santini (2002), outro cientista da computação, conclamou seus colegas de profissão a ser mais responsáveis no uso da linguagem. E adverte que:

> O irrefreável uso incorreto da linguagem em informática ameaça levar nossa profissão a se isolar da sociedade e tornar incompreensíveis nossas realizações (p. 128).

Santini concorda com o ponto que venho tentando expor:

> Outras palavras fazem mais sentido, mas estão sendo inexplicavelmente abandonadas em favor de vocábulos menos apropriados (p. 126).

Dentre os termos que ele destaca para serem desprezados estão '*data warehouse*' [armazém de dados] e '*data mart*' [mercado de dados] em vez de '*database*' [base de dados].

Uma palavra que enfrento certa dificuldade em aceitar é '*mining*' [mineração] (como em *data mining*, *text mining*, *speech mining* ou *Web mining* [mineração de dados, mineração de texto, mineração de fala ou mineração da Rede], que é amiúde usada como sinônimo de '*knowledge discovery*' [descoberta de conhecimento]. Meu pai passou muitos anos da vida numa mina de carvão do norte da Inglaterra, trabalhando como cavouqueiro. Eram longas horas de trabalho, e durante a maior parte do ano só lhe era possível ver a luz do dia uma vez por semana. Muitas vezes, cavoucava o carvão num 'veio molhado', deitado na água, de costas ou de lado, numa galeria de teto muito baixo. Não tenho certeza de que esse tipo de extração trabalhosa, na semi-escuridão, seja a analogia que os '*data miners*' [mineradores de dados] queiram realmente usar.

Minha maior queixa, porém, é o fato de o substantivo '*classification*' haver sido praticamente substituído por (pasme-se!) '*taxonomy*' (pasme-se duas vezes!!), '*ontology*' ou até (pasme-se três vezes!!!) '*taxonomized set of terms*' [conjunto taxonomizado de termos]. A maneira como estes termos são definidos em artigos recentes mostra claramente que são empregados como sinônimos de '*classification scheme*' [esquema de classificação]. Característico disso é um artigo de Hovy (2003) que define:

---

* *Monsieur* Jourdan, personagem de *Le bourgeois gentilhomme*, de Molière. (N.T.)

> [...] uma ontologia simplesmente como um conjunto taxonomizado de termos, que variam desde termos muito gerais na parte superior [...] até termos muito especializados na parte inferior (p. 48).

A 'ontologia' de Hovy torna-se uma '*concept hierarchy*' [hierarquia de conceitos] em Meng et al. (2002), que a definem como "um grande número de conceitos organizados em múltiplos níveis, de modo que os conceitos em níveis superiores possuem significados mais amplos do que os de níveis inferiores". Quando fiz o curso de biblioteconomia, zilhões de anos atrás, essas definições teriam sido definições exatas, embora muito simplistas, de classificação hierárquica.

Soergel (1999) também execrou a substituição de 'classificação' por 'ontologia' e o fez com muita propriedade:

> Uma classificação, qualquer que seja seu nome, continua sendo uma classificação. O emprego de termo diferente é sintomático da falta de comunicação entre as comunidades científicas. Ignora-se amplamente o vasto corpo de conhecimentos, que se desenvolveu em torno das classificações bibliográficas e mais geralmente da ciência da informação, sobre a estrutura das classificações e as maneiras de representá-las, bem como o imenso capital intelectual consubstanciado em muitos esquemas de classificação e tesauros. Sistemas grandes e úteis vêm sendo construídos com mais esforço do que seria necessário. Exemplos são o CYC ontology (<www.cyc.com/cyc-2-1/intro-public.html>), cuja apresentação poderia ser bastante melhorada, ou WordNet (<cogsci.princeton.edu/~wn> ou <www.notredame.ac.jp/cgi-bin/wn.cgi>), um sistema maravilhoso cuja construção teria lucrado com a aplicação da experiência com a construção de tesauros e cuja hierarquia (de conceitos) synset deveria ser tornada mais facilmente acessível com o emprego de métodos clássicos de representação de classificação. Outro exemplo é o ANSI Ad Hoc Group on Ontology Standards (<www-ksl.stanford.edu/onto-std/index.html>), que parece não contar entre seus membros com nenhum cientista da informação interessado em classificação (p. 1120).

A 'classificação' como atividade também está sendo substituída na literatura de ciência da informação pela 'categorização' (como em 'categorização de textos'), mas isso, apesar de aborrecer, não parece ser tão escandaloso.

Alguns termos da nova terminologia são superficialmente atraentes. Fui razoavelmente receptivo ao vocábulo '*summarization*' [sumarização] (porque poderia ser usado para abarcar 'abstracting' [redação de resumos], 'extracting' [extratação] e até mesmo 'annotation' [anotação]) até que descobri que um livro importante sobre o assunto (Endres-Niggemeyer, 1998) inclui a indexação temática como uma forma de sumarização. Embora um conjunto de termos de indexação possa, de fato, funcionar como uma espécie de resumo do conteúdo, a sumarização não é, com certeza, o principal objetivo da indexação.

Neste livro, sempre que possível, ative-me à terminologia antiga. Por razões de clareza, usei alguns poucos termos novos, como metadados, mas o fiz com relutância.

LISTA DE FIGURAS

Parte 1

Teoria, princípios e aplicações

CAPÍTULO 1

Introdução

O propósito principal da elaboração de índices e resumos é construir *representações* de documentos publicados numa forma que se preste a sua inclusão em algum tipo de *base de dados*. Essa base de dados de representações pode ser impressa (como numa publicação de indexação/resumos; por exemplo, o *Chemical Abstracts* ou o *Engineering Index*), em formato eletrônico (quando a base de dados muitas vezes será o equivalente aproximado de um serviço impresso), ou em fichas (como num catálogo convencional de biblioteca).

A função das operações de indexar/resumir, no âmbito maior das atividades de recuperação da informação, acha-se esquematizada na figura 1. Em primeiro lugar, o produtor da base de dados seleciona da população de documentos recém-publicados aqueles que atendam a certos critérios para sua inclusão na base de dados. O mais óbvio desses critérios é o assunto de que trata o documento. Outros critérios, no entanto, como o tipo de documento, a língua em que se acha escrito, ou sua origem, também são importantes. No caso das bases de dados que lidam principalmente com artigos de periódicos, os critérios de seleção comumente estarão centrados no periódico e não no artigo; ou seja, alguns periódicos serão incluídos e outros não (embora alguns periódicos sejam indexados em sua inteireza e outros o sejam de forma seletiva). A cobertura proporcionada por muitas bases de dados é, em grande medida, determinada por razões de custo–eficácia. Particularmente no caso de bases de dados que abranjam um campo altamente especializado, elas somente incluirão aqueles periódicos que publicam prioritariamente artigos sobre os assuntos de interesse.

Os itens selecionados para inclusão na base de dados serão 'descritos' de várias formas. Os processos de catalogação descritiva (que não aparecem na figura 1) identificam autores, títulos, fontes, e outros elementos bibliográficos; os processos de indexação identificam o assunto de que trata o documento; e o resumo serve para sintetizar o conteúdo do item. Os termos utilizados na indexação serão com freqüência extraídos de algum tipo de vocabulário controlado, como um tesauro (o 'vocabulário do sistema' da figura 1), mas, em vez disso, podem ser termos 'livres' (por exemplo, extraídos do próprio documento).*

---

* Os termos utilizados podem, genericamente, ser designados como 'termos de indexação', embora, muitas vezes, seja também empregada a palavra 'descritores', em particular quando nos estamos referindo a termos de um tesauro. Neste livro, ambas as expressões são usadas de modo equivalente.

Estas atividades de descrição criam representações dos documentos numa forma que se presta para sua inclusão na base de dados. Os próprios documentos normalmente serão destinados a um tipo diferente de base de dados (o acervo de documentos) como é o caso das estantes de uma biblioteca.

Os membros da comunidade a ser atendida utilizarão a base de dados, fundamentalmente, para satisfazer a diferentes necessidades de informação. Para lograr isso, devem converter uma necessidade de informação em algum tipo de

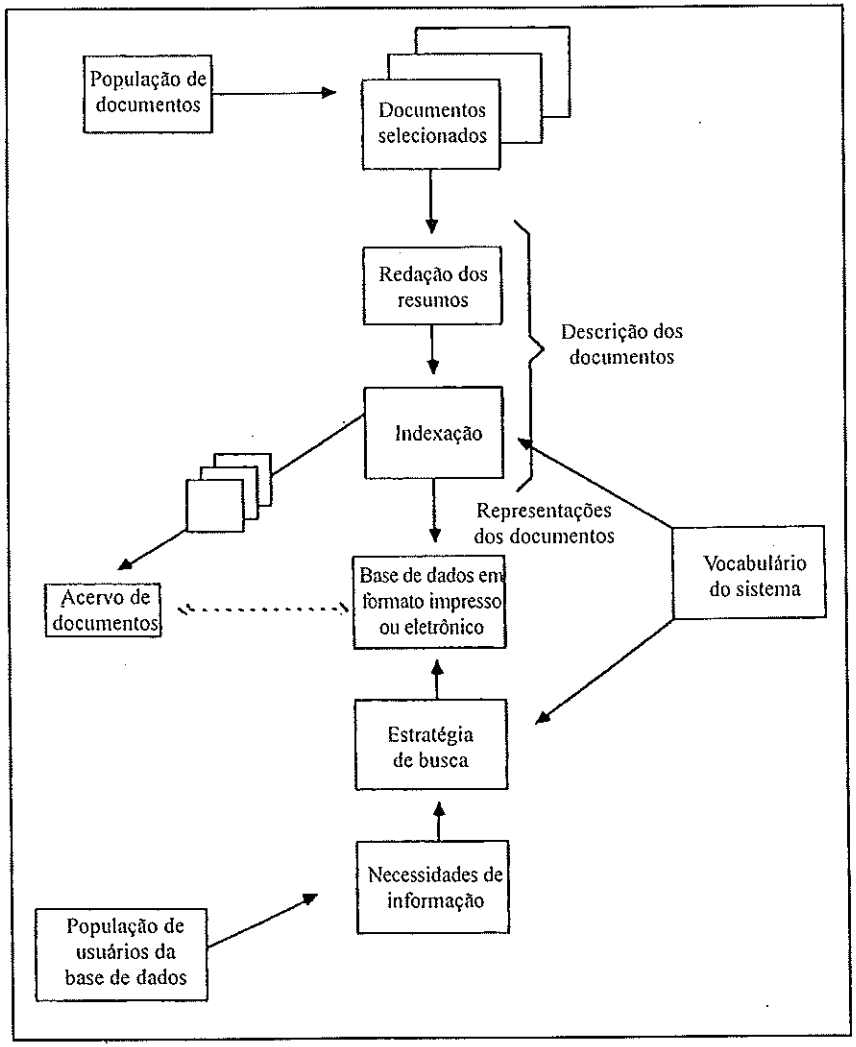


FIGURA 1
A função da elaboração de índices e resumos no quadro mais amplo da recuperação da informação

'estratégia de busca', a qual pode ser tão simples quanto a escolha de um único termo para consultar um índice impresso ou um catálogo em fichas, ou exigir a combinação de muitos termos numa estratégia mais elaborada e complexa, empregada para consultar uma base de dados mantida localmente ou conectada em linha a alguma rede de computadores.

O que se almeja, evidentemente, ao fazer uma busca numa base de dados, é encontrar documentos que sejam úteis para satisfazer a uma necessidade de informação, e evitar a recuperação de itens inúteis. 'Relevante' e 'pertinente' são termos freqüentemente empregados para se referir a itens 'úteis', e foram definidos de diferentes formas. Há muito desacordo sobre o que realmente significam 'relevância' e 'pertinência' (Lancaster e Warner, 1993). Neste livro considerarei como sinônimas as expressões 'útil', 'pertinente' e 'relevante para uma necessidade de informação'. Ou seja, um documento pertinente (útil) é aquele que contribui para satisfazer a uma necessidade de informação.

O problema da recuperação da informação está representado graficamente na figura 2. O retângulo inteiro representa uma base de dados e os itens que contém. Os itens com sinal de adição (+) são aqueles que um consulente hipotético consideraria úteis para atender a uma necessidade de informação atual, e os itens com sinal de subtração (–) são aqueles que não consideraria úteis. Para qualquer necessidade específica de informação haverá muito mais itens – do que itens +. Na realidade, se se desenhasse o diagrama 'em escala',

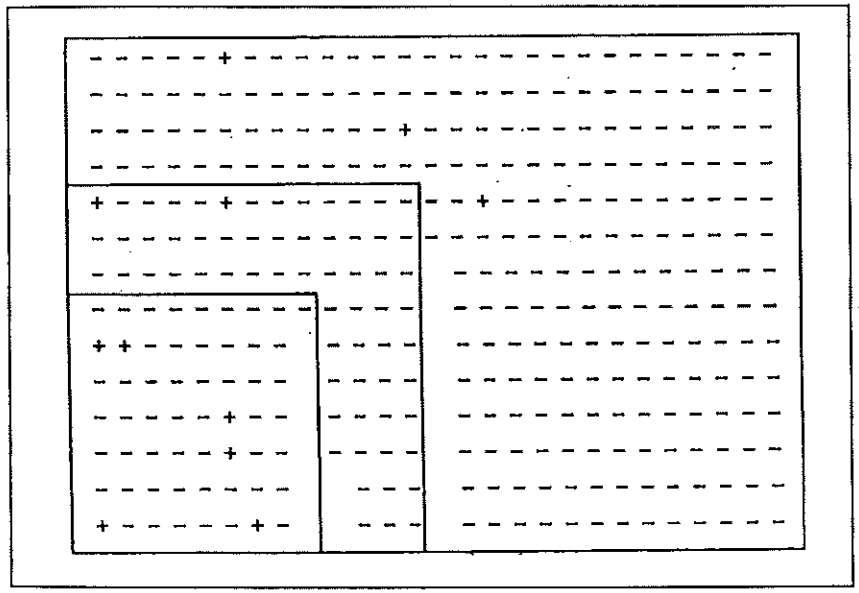


FIGURA 2
O problema da recuperação de itens pertinentes de uma base de dados

seria quase certo que os onze itens úteis estariam acompanhados de toda uma muralha de itens inúteis. O problema está em recuperar tantos itens úteis quantos for possível, e o menor número possível de itens inúteis.

O menor dos dois retângulos internos da figura 2 representa os resultados de uma busca realizada na base de dados, que recuperou 57 itens, seis dos quais foram úteis e 51 inúteis. A relação entre itens úteis e o total de itens recuperados (6/57 ou cerca de 10% neste caso) é comumente denominada *coeficiente de precisão.* O índice empregado habitualmente para expressar a extensão com que todos os itens úteis são encontrados é o *coeficiente de revocação.* No presente exemplo, o coeficiente de revocação é de 6/11 ou cerca de 54%.

Nessa situação, provavelmente seria preciso, para melhorar a revocação, fazer uma busca mais genérica. Essa busca é representada pelo maior dos dois retângulos internos. Ao fazer a busca de modo mais genérico, aumentou-se a revocação para 8/11 (73%), mas a precisão caiu ainda mais para 8/112, ou cerca de 7%. Uma característica lamentável, inerente à recuperação da informação, é que uma melhoria da revocação em geral implica perda de precisão e vice-versa.

A figura 2 sugere outro fenômeno. Talvez fosse possível fazer uma busca suficientemente genérica para localizar todos os itens úteis (isto é, alcançar 100% de revocação); entretanto a precisão seria provavelmente inaceitável. Ademais, quanto maior for a base de dados, menos aceitável será uma baixa precisão. Embora o usuário esteja disposto a examinar, por hipótese, 57 itens, a fim de encontrar seis que lhe sejam úteis, talvez se sinta muito menos inclinado a examinar 570 resumos para encontrar 60 que sejam úteis. Em bases de dados muito grandes torna-se, portanto, progressivamente mais difícil alcançar um nível de revocação aceitável com um nível de precisão satisfatório, uma situação que chegou a um ponto crítico quando se procura informação na internet.

Neste livro emprego o termo *revocação* [*recall*] para designar a capacidade de recuperar documentos úteis, e *precisão* para designar a capacidade de evitar documentos inúteis. Existem outras medidas do desempenho para buscas realizadas em bases de dados (ver, por exemplo, Robertson, 1969), algumas das quais são matematicamente mais exatas, porém a *revocação* e a *precisão* compõem o quadro geral e ainda parecem ser as medidas óbvias a serem utilizadas para expressar os resultados de qualquer busca que simplesmente divida uma base de dados em duas partes (recuperados e não recuperados).* A figura 1 deixa evidente que são muitos os fatores que determinam se uma busca numa base de dados é ou não bem sucedida. Entre tais fatores encontra-se a cobertura da base de dados, sua política de indexação, sua prática de indexação, sua política e prática de redação de resumos, a qualidade do vocabulário empregado na indexação, a qualidade das estratégias de busca, e assim por diante. Este livro não

* Uma busca que classifique os resultados em ordem de 'relevância provável' exige uma medida um tanto diferente, a qual, com efeito, compara a classificação [*ranking*] obtida com uma classificação ideal.

enceta qualquer esforço no sentido de tratar de todos esses fatores (ainda que todos estejam inter-relacionados), mas se concentra nas atividades importantes de descrição do documento ou, pelo menos, aquelas que dizem respeito ao conteúdo dos documentos.

Em princípio, a base de dados representada na figura 1 poderia ser a totalidade do conteúdo da Rede Mundial (World Wide Web) (doravante denominada simplesmente a Rede). No entanto, o diagrama não representa a situação da Rede tão bem quanto representa bases de dados, como o catálogo de uma biblioteca universitária ou uma base de dados de registros bibliográficos de artigos de periódicos, como a base de dados MEDLINE da National Library of Medicine. Uma vez que qualquer organização ou qualquer pessoa pode criar uma página na Rede, não está em causa nenhum processo de seleção real. Ademais, embora os sítios da Rede possam incluir algum tipo de dado descritivo sobre seu conteúdo (normalmente denominados 'metadados'; ver a nota que precede imediatamente este capítulo), muitos não o fazem, e os dados descritivos são parte integrante das próprias páginas da Rede, não se encontrando numa base de dados separada. Além do que, a indexação e a elaboração de resumos de conteúdos da Rede por seres humanos constituem mais a exceção do que a regra, de modo que a maior parte das buscas ali feitas ocorre no texto integral dos sítios acessados por determinado mecanismo de busca. Nos casos em que são realizadas operações de indexação ou resumo, o provável é que sejam efetuadas 'automaticamente' por meio de várias etapas de processamento informatizado. Esses procedimentos automáticos, junto com as buscas em textos completos e o caso específico da Rede, são tratados nos capítulos finais deste livro. Embora a figura 1 não corresponda exatamente à situação encontrada na Rede, a figura 2, sim. Isto é, o problema de busca mostrado ali é igualmente pertinente às buscas feitas na Rede, exceto que multiplicada por ordens de grandeza.

CAPÍTULO 2

Princípios da indexação

Embora o título deste livro se refira à 'indexação', seu alcance limita-se, de fato, à indexação de assuntos e redação de resumos. A indexação de assuntos e a redação de resumos são atividades intimamente relacionadas, pois ambas implicam a preparação de uma *representação* do conteúdo temático dos documentos. O resumidor redige uma descrição narrativa ou síntese do documento, e o indexador descreve seu conteúdo ao empregar um ou vários termos de indexação, comumente selecionados de algum tipo de vocabulário controlado.

O principal objetivo do resumo é indicar de que trata o documento ou sintetizar seu conteúdo. Um grupo de termos de indexação serve ao mesmo propósito. Por exemplo, o seguinte conjunto de termos proporciona uma idéia bastante razoável sobre os assuntos tratados num relatório hipotético:

Centros de Informação
Compartilhamento de Recursos
Catálogos Coletivos
Catalogação Cooperativa
Redes em Linha
Empréstimos entre Bibliotecas

Em certo sentido, essa lista de termos pode ser vista como uma espécie de minirresumo. Serviria a tal propósito se todos os termos fossem reunidos num índice publicado, copiados pela impressora ou mostrados na tela para representar um item recuperado numa base de dados, como resultado de uma busca em linha.

De modo mais evidente, os termos atribuídos pelo indexador servem como pontos de acesso mediante os quais um item é localizado e recuperado, durante uma busca por assunto num índice publicado ou numa base de dados eletrônica.* Assim, num índice impresso, convém que se possa encontrar o item hipotético mencionado anteriormente sob qualquer um dos seis termos. Num sistema de recuperação informatizado, evidentemente, seria natural encontrá-lo sob qualquer um desses termos ou, de fato, sob qualquer combinação deles.

* Outros autores empregam terminologia diferente para designar a indexação e os termos de indexação sem que isso altere de modo relevante o significado adotado neste livro. Por exemplo, Anderson (1985) vê os termos como 'indicadores' de conteúdo; indexação como "o processo de indicar o conteúdo e características afins de um documento". O'Connor (1996) prefere o termo 'apontamento' [*pointing*]: os termos de indexação são apontadores; indexação é a tarefa de atribuir apontadores úteis a fontes de informação.

A diferença entre indexação e redação de resumos está se tornando cada vez mais difusa. Por um lado, uma lista de termos de indexação pode ser copiada pela impressora ou mostrada na tela de modo a constituir um minirresumo. Por outro lado, o texto de resumos pode ser armazenado num sistema informatizado de modo a permitir a realização de buscas por meio da combinação de palavras que ocorram nos textos. Esses resumos podem ser utilizados no lugar de termos de indexação, permitindo o acesso aos itens, ou complementar os pontos de acesso proporcionados pelos termos de indexação. Em certa medida isso modifica a função do resumidor, que deve agora preocupar-se não só em redigir uma descrição clara e de boa qualidade do conteúdo do documento, mas também em criar um registro que seja uma representação eficaz para fins de recuperação.

Se a indexação e a redação de resumos fossem consideradas como atividades inteiramente complementares, a natureza da atividade de indexação sofreria algum tipo de mudança. Por exemplo, o indexador se concentraria na atribuição de termos que complementassem os pontos de acesso existentes no resumo. Tal complementaridade, porém, deve ser inteiramente reconhecida e compreendida pelo usuário da base de dados. Do contrário, um conjunto de termos de indexação isolados daria uma imagem bastante equivocada do conteúdo de um item.

Extensão do registro

Uma das propriedades mais importantes de uma representação de conteúdo temático é sua extensão. O efeito da extensão do registro acha-se exemplificado na figura 3. No lado esquerdo da figura, encontram-se várias representações do conteúdo de um artigo de periódico na forma de texto narrativo; no lado direito, estão duas representações na forma de listas de termos de indexação.

O título contém uma indicação geral sobre aquilo de que trata o artigo. O resumo breve oferece mais detalhes, indicando que o artigo apresenta resultados da pesquisa e identificando as principais questões analisadas. O resumo ampliado vai mais além, identificando todas as questões focalizadas na pesquisa e informando sobre o tamanho da amostra utilizada no estudo.

Quanto mais informações são apresentadas, mais claramente a representação revela o alcance do artigo, tornando-se mais provável que venha a indicar para o leitor se esse artigo satisfaz ou não a uma necessidade de informação. Por exemplo, alguém talvez esteja à procura de artigos que mencionem as atitudes norte-americanas em relação a vários líderes árabes. O título não traz indicação alguma de que esse tópico específico seja analisado, e o resumo breve, ao focalizar outros tópicos, sugere que talvez isso não aconteça. É somente o resumo ampliado que mostra que o artigo inclui informações sobre esse assunto.

Também, quanto maior a representação, mais pontos de acesso ela proporciona. Se as palavras do título fossem os únicos pontos de acesso, esse item provavelmente não seria localizado em muitas buscas para as quais poderia ser considerado uma resposta válida. À medida que se aumenta a extensão da repre-

sentação também se aumenta a recuperabilidade do item. É provável que somente com o resumo ampliado fosse possível recuperar esse item durante uma busca de informações sobre as atitudes norte-americanas em face dos líderes árabes.

Título

Pesquisa nacional de opinião pública sobre as atitudes norte-americanas a respeito do Oriente Médio

Resumo (breve)

Uma pesquisa realizada por telefone em 1985 apresenta opiniões sobre tópicos como: a ajuda norte-americana a Israel e ao Egito; se os EUA devem tomar o partido de Israel, das nações árabes, ou de nenhum destes; se a OLP deve participar de uma conferência de paz; e se um Estado palestino independente é um pré-requisito para a paz.

Resumo (ampliado)

Em 1985 foram feitas entrevistas por telefone com uma amostra probabilística de 655 norte-americanos. Obtiveram-se respostas às seguintes questões: o estabelecimento de um Estado palestino é essencial para a paz; deve ser reduzida a ajuda norte-americana a Israel e ao Egito; os EUA devem participar de uma conferência de paz que inclua a OLP; os EUA não devem favorecer nem Israel nem as nações árabes, mas, sim, manter relações amistosas com eles? Também se expressaram opiniões sobre os principais líderes do Oriente Médio (Hussein, Arafat, Peres, Mubarak, Fahd, Assad), especialmente seus esforços pela paz, e se os entrevistados achavam que possuíam ou não informações suficientes sobre os diversos grupos nacionais da região.

Indexação (seletiva)

OPINIÃO PÚBLICA
PESQUISAS POR TELEFONE
ESTADOS UNIDOS
ATITUDES
ORIENTE MÉDIO

Indexação (exaustiva)

OPINIÃO PÚBLICA
PESQUISAS POR TELEFONE
ESTADOS UNIDOS
ATITUDES
ORIENTE MÉDIO
ISRAEL
EGITO
NAÇÕES ÁRABES
ORGANIZAÇÃO PARA A LIBERTAÇÃO DA PALESTINA
CONFERÊNCIAS DE PAZ
PAZ
ESTADO PALESTINO
AJUDA EXTERNA
LÍDERES POLÍTICOS

FIGURA 3
Efeito da extensão do registro sobre a recuperabilidade

A mesma situação aplica-se à indexação. A indexação seletiva, que inclua apenas cinco termos, apresenta uma indicação muito geral daquilo de que trata o artigo (aproximadamente equivalente, neste caso, ao título) e um nível de acesso muito limitado. A indexação mais exaustiva proporciona uma indicação muito melhor do assunto específico de que trata o artigo, bem como possibilita muito mais pontos de acesso.

Etapas da indexação de assuntos

A indexação de assuntos envolve duas etapas principais:

1. Análise conceitual, e
2. Tradução.

Intelectualmente são etapas totalmente distintas, embora nem sempre sejam diferençadas com clareza e possam, de fato, ocorrer de modo simultâneo.

A análise conceitual, em primeiro lugar, implica decidir do que trata um documento — isto é, qual o seu assunto. Os termos que aparecem na lista à direita, na figura 3, representam a análise conceitual de um artigo feita por este autor — aquilo que, segundo sua opinião, constituía o assunto do artigo.

Esta afirmativa sobre análise conceitual está bastante simplificada. A indexação de assuntos é normalmente feita visando a atender às necessidades de determinada clientela — os usuários de um centro de informação ou de uma publicação específica. Uma indexação de assuntos eficiente implica que se tome uma decisão não somente quanto ao que é tratado num documento, mas também por que ele se reveste de provável interesse para determinado grupo de usuários. Em outras palavras, não há um conjunto 'correto' de termos de indexação para documento algum. A mesma publicação será indexada de modo bastante diferente em diferentes centros de informação, e deve ser indexada de modo diferente, se os grupos de usuários estiverem interessados no documento por diferentes razões.*

O indexador, então, deve formular várias perguntas sobre um documento:

1. De que trata?
2. Por que foi incorporado a nosso acervo?
3. Quais de seus aspectos serão de interesse para nossos usuários?

Esta situação acha-se bem exemplificada na figura 4. Esse exemplo hipotético refere-se a relatório publicado pela National Aeronautics and Space Administration (NASA) a respeito de um vôo espacial tripulado. Ao incorporar esse relatório à sua própria base de dados, a NASA provavelmente estará interessada em todas as suas facetas e o indexará exaustivamente, procurando abranger todos os seus aspectos, talvez em nível razoavelmente genérico. Uma parte do relatório refere-se ao traje usado pelos astronautas, mencionando alguns compostos novos de borracha sintética empregados em partes desse traje. Isto faz com que o relatório seja interessante para uma fábrica de borracha. Ao ser incorporado ao acervo de documentos dessa fábrica, o relatório será indexado, porém, de modo bastante diferente. Serão usados termos altamente específicos para indexar os compostos novos, e o termo genérico TRAJES ESPACIAIS talvez seja empregado para indicar determinada aplicação para esses compostos. Uma empresa metalúrgica poderá interessar-se pelo mesmo relatório por um motivo diferente: ele menciona uma nova técnica de soldagem desenvolvida

* Dabney (1986a) admitiu isso ao fazer uma distinção entre indexação orientada para o documento e indexação orientada para a consulta. Acha-se também implícito no método chamado '*gedanken*' proposto por Cooper (1978).

para unir certas ligas na construção do veículo espacial. Neste caso, será indexado sob os termos relativos a soldagem, os termos apropriados para metais e talvez o termo de aplicação genérica VEÍCULOS ESPACIAIS. A fábrica de borracha indexa o relatório de forma muito diferente daquela adotada pela empresa metalúrgica, e nenhum desses conjuntos de termos se assemelha à lista mais exaustiva adotada pela própria NASA.

E assim é porque assim tem que ser. Quanto mais especializada a clientela de um centro de informação maior a probabilidade de que a indexação possa e deva ser feita sob medida, ajustando-se com precisão aos interesses do grupo. Somente entre instituições de caráter mais genérico, como, por exemplo, bibliotecas universitárias gerais, é que existe a possibilidade de uma delas indexar um documento exatamente da mesma forma que outra. Fidel (1994) emprega a expressão 'indexação centrada no usuário' para designar o princípio da indexação que se baseia nos pedidos que são esperados de determinada clientela.

Relatório Técnico da NASA Com a Descrição de uma Nova Missão Espacial Tripulada

NASA	Fábrica de Borracha	Empresa Metalúrgica
____ Indexação ____ exaustiva ____ abrangendo ____ todos os ____ aspectos ____ num nível ____ um tanto ____ genérico ____ ____ ____	____ Novos ____ compostos ____ de ____ borracha ____ sintética ____ TRAJES ESPACIAIS	____ Novas ____ técnicas ____ de ____ soldagem e ____ metais ____ envolvidos ____ VEÍCULOS ESPACIAIS

FIGURA 4
Exemplo de um documento indexado segundo diferentes pontos de vista

Hjørland (2001) concorda que a indexação deve ser moldada para se ajustar às necessidades de determinada clientela:

> Uma vez que qualquer documento pode, em princípio, proporcionar respostas a uma infinidade de questões, as análises de assuntos devem estabelecer prioridades baseadas nos grupos de usuários específicos atendidos (ou serviços específicos proporcionados na ecologia da informação). O assunto de um documento é assim relativo ao objetivo do serviço de informação específico. Defino assunto [...] como os *potenciais* epistemológicos ou informativos dos documentos. A melhor análise temática é a que faz o melhor prognóstico quanto ao uso futuro do documento (p. 776).

Este aspecto foi também examinado por Bates (1998):

> [...] o desafio para o indexador é tentar antecipar quais os termos que as pessoas que possuem lacunas de informação de vários tipos procurariam nos casos em que o registro de que dispõem, de fato, fica a meio caminho de satisfazer a necessidade de informação do usuário. Quando se pensa em tal desafio, é possível perceber que se trata de algo muito peculiar. Quais os tipos de necessidades de informação que as pessoas teriam e as levariam a querer informações que o registro, de fato, contém? (p. 1187).

Mai (2001), que se vale da semiótica na análise do processo de indexação temática, faz uma descrição lúcida das dificuldades que caracterizam o esforço de reconhecer por que determinado documento viria a ter interesse para futuros usuários:

> Seria quase impossível, naturalmente, para qualquer pessoa ou, neste caso, qualquer indexador, precisar *todas* as idéias e significados que estivessem associados a qualquer documento, posto que sempre haverá idéias e significados potenciais que diferentes pessoas em diferentes momentos e lugares poderão descobrir nesse documento. Além do que, seria quase impossível prever com exatidão quais das inúmeras idéias e significados que estivessem associados ao documento seriam especificamente úteis para os usuários ou dariam ao documento alguma utilidade duradoura. É da máxima importância reconhecer e aceitar essa indefinição fundamental. O indexador deve compreender, desde o início, que jamais descobrirá todas as idéias e significados que estariam associados ao documento e que, portanto, não é possível descrever todas essas idéias e significados (p. 606).

Layne (2002), referindo-se especificamente à indexação de imagens no campo da arte, também admite a necessidade de indexação diferente, com terminologia diferente, para públicos diferentes:

> O segundo aspecto da escolha do vocabulário para imagens artísticas está em que uma imagem pode interessar a várias disciplinas com diferentes vocabulários. Por exemplo, *O nascimento de Esaú e Jacó* teria interesse para historiadores da medicina que desejassem usar o vocabulário médico, ao invés de um vocabulário mais comum, na busca de imagens. Evidentemente, não é prático empregar todos os vocabulários possíveis quando da criação de acessos temáticos para imagens artísticas. Mas, se se souber ou houver a intenção de que determinado acervo de imagens artísticas será usado por determinada disciplina, talvez valha a pena ponderar quanto ao uso de um vocabulário especializado além do vocabulário geral. Por exemplo, uma imagem de tulipas seria indexada como 'tulipas' ou mesmo 'flores' para usuários comuns, e os nomes científicos das espécies, como *Tulipa turkestanica*, seriam empregados como termos de indexação, caso houvesse botânicos entre os usuários pretendidos (p. 15).

Para certos tipos de materiais, a indexação orientada para o usuário pode até ser mais importante do que o é no caso de artigos de periódicos, livros ou relatórios técnicos. Por exemplo, certos autores, como Shatford (1986) e Enser (1995), salientam que diferentes grupos de usuários podem ver os acervos de imagens de modo bastante diferente. O que levou Brown et al. (1996) a sugerirem a necessidade de um tratamento 'democrático' da indexação, em que os usuários acrescentariam aos registros termos de sua própria escolha, quando isso fosse necessário e apropriado.

Hidderley e Rafferty (1997) apresentam um método de tratamento democrático da indexação. Uma amostra de usuários recebe um objeto (livro, artigo, imagem) junto com uma indexação desse objeto que reflete uma 'visão pública' (por exemplo, um conjunto de termos extraídos de um tesauro por indexadores experientes). Os usuários alteram a visão pública de modo a refletir sua própria 'visão particular'. Com base em múltiplas visões particulares de um conjunto de objetos, surge uma nova visão pública. Adota-se um processo de harmonização para chegar à visão pública final. Esse processo leva em conta quantos usuários associaram determinado termo a determinado objeto. Em especial, os autores defendem um tratamento 'democrático' da indexação de obras de ficção porque, como salientam, "o texto ficcional pode ser lido de muitas maneiras diferentes".

Métodos colaborativos ou 'democráticos' são, no mais das vezes, recomendados para o caso da indexação de imagens (ver o capítulo 13).

Os métodos colaborativos de indexação são, obviamente, mais viáveis em ambiente de biblioteca digital. Isto é, usuários de uma biblioteca podem oferecer novos termos de indexação aos itens que consultam, e esses termos fornecidos pelos usuários serão então armazenados em novo campo do registro. Villarroel et al. (2002) propõem um enfoque em que os usuários destacam seções do texto digital que consideram importantes, e esse destaque pode levar à revisão dos 'pesos' (ver capítulo 11) relativos aos termos de indexação ou palavras do texto.

Há uma importante lição a tirar dos princípios da indexação orientada para o usuário. É preciso que os indexadores saibam muito mais do que os princípios da indexação. Devem, em especial, estar inteiramente a par dos interesses da comunidade atendida e das necessidades de informação de seus membros. Na realidade, recomenda-se, usualmente, que o indexador não fique 'nos bastidores', mas que também procure desempenhar outras atividades, inclusive a de bibliotecário de referência, onde participam de buscas nos registros que criaram.

Pode-se avançar ainda mais com o princípio da indexação orientada para o usuário ao sustentar que, em relação a determinado acervo de documentos e determinado grupo de usuários, qualquer conjunto ideal de termos de indexação será ideal *somente em determinado ponto no tempo*. Passados alguns anos, o mesmo grupo de usuários poderá precisar de acesso ao mesmo acervo (ou outro bastante semelhante) a partir de perspectivas diferentes. Um exemplo óbvio seria uma coleção de relatórios técnicos dentro de uma instituição de pesquisa: as mudanças de prioridades e os interesses de pesquisa da instituição podem alterar a forma como a coleção é útil para a comunidade. Isso pode ser verdade, em especial, no caso de pesquisas interdisciplinares. De fato, pode-se alegar que, num mundo ideal, um acervo seria organizado (isto é, indexado) em torno dos interesses de determinado projeto de pesquisa. Quando o projeto mudasse, o acervo seria reorganizado em torno das novas exigências. Naturalmente, o custo da reindexação e reorganização em geral torna essa proposta economicamente pouco atraente. Weinberg (1992) salientou a impermanência do acesso temático

e o fato de ser 'relativo'. No entanto, ela baseia essa afirmação no fato de que os vocabulários (por exemplo, cabeçalhos de assuntos, classificações) mudam e não no fato de que as necessidades e interesses dos usuários mudam.

Mai (2000) também adverte que a indexação orientada para o usuário somente pode estar voltada para determinado conjunto de usuários em determinado ponto no tempo:

> Se se focalizar exclusivamente o aspecto da representação, ignorando os usuários futuros, corre-se o risco de representar os documentos de uma forma que não terá qualquer serventia para os usuários. Um indexador que não dê muita atenção aos usuários poderá optar por representar assuntos de documentos que não tenham interesse para eles, ou usar um vocabulário diferente do vocabulário deles, ou representar o assunto em nível que seja muito genérico ou muito específico para eles. No entanto, se o indexador der excessiva atenção aos usuários do sistema, poderá representar os documentos numa forma tal que a representação temática dos documentos somente atenda aos usuários atuais e às necessidades de informação atuais (p. 294).

Atinência

Nas considerações anteriores não se fez qualquer tentativa para definir a expressão 'de que trata um documento': a expressão 'de que trata' era simplesmente um sinônimo para 'tem por assunto'. Ou seja, usou-se 'de que trata um documento' para designar o mesmo que 'os assuntos de um documento'. Estas expressões talvez não sejam muito precisas e não é fácil definir 'trata de' e 'tem por assunto'. Apesar disso, são expressões que soam aceitáveis para a maioria das pessoas, sendo por elas compreendidas. Não pretendo partir para uma discussão filosófica sobre o significado de 'trata de' ou 'atinência'.* Vários autores já o fizeram. E nem assim conseguiram esclarecer a situação, pelo menos no que tange à atividade da indexação de assuntos. Beghtol (1986) e Hutchins (1978) recorrem ambos à lingüística do texto ao examinar esta questão; Maron (1977) adota um enfoque probabilístico, e Swift et al. (1978) são cautelosos ao salientar que a atinência na indexação talvez não coincida com a atinência que as pessoas que estão em busca de informações têm em mente. Wilson (1968) chega ao ponto de sugerir que a indexação de assuntos se defronta com problemas 'intratáveis', visto ser tão difícil decidir do que trata um documento.

Moens et al. (1999) afirmam que um texto não possui uma 'atinência' intrínseca, mas que também possui diferentes 'significados' de acordo com "o uso particular que uma pessoa pode fazer da atinência em dado momento".

Layne (2002) faz distinção entre 'de-ência' ['*of-ness*'] e atinência ['*aboutness*'] no caso de imagens artísticas:

---

* O autor emprega os termos ingleses *about* e *aboutness*. O primeiro traduzimos por 'trata de' e o segundo por 'atinência'. Outros traduzem *aboutness* por 'tematicidade', 'temática', 'acerca-de', 'ser acerca-de', 'ser sobre algo', etc. (N.T.)

> Menos óbvio do que a *de-ência* [*of-ness*] de uma obra de arte, mas muitas vezes mais instigante, é aquilo *de que trata* a obra de arte. [...] Às vezes, a atinência [*about-ness*] de uma obra de arte é relativamente óbvia, como na *Alegoria da justiça*, de Georg Pencz. [...] Essa é a imagem *de* [*of*] uma mulher despida que segura uma espada e uma balança, mas o título nos diz que a imagem é uma figura alegórica que representa a justiça ou, em outras palavras, que a imagem *trata do* [*is about*] conceito abstrato de 'justiça'. No desenho de Goya *Despreciar los insultos* [...] a *atinência* é um pouco menos óbvia, mas é claro que essa obra possui algum significado além simplesmente do que mostra *de*. De fato, uma descrição do que contém — um homem, talvez o próprio Goya, gesticulando para dois anões uniformizados — não basta realmente para dar sentido à imagem; ela simboliza algo mais, *trata de* algo mais: a relação entre Espanha e França no início do século XIX ou, mais especificamente, a atitude pessoal de Goya em relação à ocupação da Espanha pela França (p. 4).

Ela acredita que essa distinção é válida e que, na recuperação, deveria ser possível separar uma da outra:

> [...] possibilita recuperar, por exemplo, exatamente aquelas imagens que sejam *da* 'morte' e excluir as que *tratam da* 'morte'. Também permite a subdivisão de grandes conjuntos de imagens recuperadas com base nessas distinções. Por exemplo, uma pesquisa sobre 'morte' como assunto recuperaria imagens subdivididas em grupos baseados em se a imagem representa explicitamente a 'morte' ou se *trata do* tema da 'morte' (p. 13).

Bruza et al. (2000) focalizam a atinência de uma perspectiva lógica. Tentam "formalizar a relevância lógica mediante a formalização de propriedades do senso comum que descrevem a relação de atinência". Também trabalham com a 'não-atinência' e a interação entre atinência e não-atinência. No contexto da recuperação da informação, a não-atinência constitui realmente uma situação mais simples porque a grande maioria dos itens em qualquer base de dados evidentemente não guarda qualquer relação possível com qualquer consulta ou necessidade de informação (isto é, são naturalmente itens 'não-atinentes').

O tema da atinência está relacionado muito de perto com o da relevância — isto é, a relação entre um documento e uma necessidade de informação ou entre um documento e um enunciado de necessidade de informação (uma consulta). O tema da relevância/pertinência produziu um grande volume de debates e publicações. Encontra-se em Mizzaro (1998) um apanhado muito completo. Hjørland (2000) salienta que a relevância é dependente dos pressupostos teóricos que orientam o comportamento da pessoa que busca informação.

Conforme Harter (1992) ressaltou, no entanto, um documento pode ser relevante para uma necessidade de informação sem 'tratar' dessa necessidade de informação. Por exemplo, se escrevo sobre o tema das barreiras à comunicação, uma história do latim talvez tenha alguma relevância, principalmente se lidar com a utilização atual do latim pela Igreja Católica e com as instituições que hoje em dia se esforçam para promover seu uso mais amplo. Não obstante, ainda que possa inspirar-me nessa fonte ao escrever meu artigo, poucas pessoas alegariam

que ele 'trata' da comunicação internacional, sendo improvável que venha a ser indexado desta forma, a menos que o autor faça menção explicitamente ao aspecto da comunicação internacional.

Wong et al. (2001) tratam 'atinência' como sendo mais ou menos sinônimo de 'relevância':

> [...] se um dado documento D *trata do* pedido Q, então existe uma alta probabilidade de que D será relevante em relação à necessidade de informação associada. Assim, o problema da recuperação da informação se reduz à decisão acerca da relação de atinência entre documentos e pedidos (p. 338).

Eles relacionam a atinência diretamente às medidas de revocação e precisão.

Continuam a aparecer na literatura artigos sobre atinência. Hjørland (2001) e Bruza et al. (2000) são exemplos. Embora possam apresentar algum interesse acadêmico (Hjørland dá-se ao trabalho de tentar diferençar termos como 'assunto', 'tópico', 'tema', 'domínio', 'campo' e 'conteúdo'), não têm qualquer importância prática para o indexador, que fará bem se ignorar essas diferenças semânticas e simplesmente atribuir ao item os rótulos que o tornarão utilmente recuperável pelos membros de uma comunidade-alvo.

Em outras palavras, será que precisamos realmente compreender o que é 'atinência' a fim de indexar de maneira eficiente? Não bastará que sejamos capazes de reconhecer que um documento tem interesse para determinada comunidade pelo fato de contribuir para nossa compreensão dos tópicos *X*, *Y* e *Z*? O reconhecimento de que realmente contribui para isso exemplifica o processo que chamamos 'análise conceitual', enquanto o processo de 'tradução' envolve uma decisão sobre quais dos rótulos disponíveis melhor representam *X*, *Y* e *Z*. 'Conceito' é outra palavra sobre a qual alguns autores gostam de filosofar (ver, por exemplo, Dahlberg [1979]). Neste livro emprego-a para referir-me a um assunto estudado por um autor ou representado de alguma outra forma (por exemplo, numa fotografia ou outra imagem). 'Análise conceitual', portanto, significa nada mais do que a identificação dos assuntos estudados ou representados num documento. Preschel (1972) adota uma abordagem muito prática. Para ela, 'conceito' significa 'matéria indexável', e 'análise conceitual' é a 'percepção pelo indexador de matéria indexável'. Tinker também adota uma posição prática (1966):

> Ao atribuir um descritor [isto é, um termo de indexação] a um documento, o indexador declara que tal descritor possui alto grau de relevância para o conteúdo do documento; quer dizer, ele declara que o significado do descritor está fortemente associado a um conceito incorporado ao documento, e que é adequado à área temática do documento (p. 97).

Wooster (1964) é ainda mais pragmático, ao se referir à indexação como a atribuição de termos "provavelmente relacionados de alguma forma com o conteúdo intelectual do documento original, para ajudar você a encontrá-lo quando precisar".

Não vejo nada de errado nessas definições ou descrições pragmáticas da indexação temática. Os puristas sem dúvida tergiversarão sobre elas argumentando que expressões como 'matéria indexável', 'relevância', 'significado', 'associado a', 'conceito', 'adequado a', 'relacionado com' e 'conteúdo intelectual' não se acham definidas precisamente de modo a satisfazer a todos. No entanto, se tivermos de chegar a um acordo quanto à definição exata dos termos antes de encetar qualquer tarefa, é improvável que cheguemos muito longe, seja na indexação seja em qualquer outra atividade.

Weinberg (1988) levanta a hipótese de que a indexação frustra o pesquisador porque ela lida apenas de forma genérica com aquilo de que 'trata' um documento e não focaliza aquilo que ele proporciona de 'novidade' a respeito do tópico. Ela afirma que esta distinção se reflete na diferença entre 'atinência' e 'aspecto', entre 'tópico' e 'comentário' ou entre 'tema' e 'rema'. Ela não consegue convencer que essas distinções sejam realmente úteis no contexto da indexação ou que seja possível para os indexadores sustentar essas distinções.

Swift et al. (1978) examinam as limitações de um enfoque baseado na atinência na indexação em ciências sociais, e recomendam que os documentos sejam indexados de acordo com os 'problemas' com os quais pareçam estar relacionados. É difícil perceber como a distinção que fazem difere da distinção, feita anteriormente neste capítulo, entre do que trata um documento e por que um determinado usuário ou grupo de usuários teria interesse nele. Crowe (1986) afirma que o indexador deve remeter ao 'ponto de vista subjetivo' do autor. Um de seus exemplos trata do tópico da depressão, o qual pode ser estudado em livros ou artigos a partir de diferentes pontos de vista (por exemplo, tratamento por meio de psicoterapia, por meio de medicamentos, etc.). Outra vez torna-se difícil vislumbrar como isso difere da prática habitual da indexação, como, por exemplo, o emprego de subcabeçalhos pela National Library of Medicine.

Breton (1981) alega que os engenheiros pouco recorrem às bases de dados porque os indexadores rotulam os documentos com os *nomes* de materiais ou dispositivos, enquanto é mais provável que os engenheiros precisem fazer as buscas a partir dos *atributos* ou das *funções* desempenhadas por esses materiais ou dispositivos. Em outras palavras, eles gostariam de localizar um material ou dispositivo que satisfizesse a algum requisito atual (quanto à resistência, condutividade, resistência à corrosão, ou coisa que o valha) sem terem de nomeá-lo. Isso não constitui uma condenação da indexação de assuntos de per si, mas das políticas de indexação adotadas pela maioria dos produtores de bases de dados. Se se diz que um novo material ou uma liga descrita num relatório possui certa resistência à tração, esta propriedade pode ser indexada (por exemplo, atribuindo o termo RESISTÊNCIA À TRAÇÃO), porém o *valor* específico dessa propriedade (isto é, a resistência alcançável) não seria indexado pela maioria dos produtores de bases de dados, embora se possa mencioná-lo no resumo. Naturalmente, não há razão para que os valores não sejam indexados (por exemplo, o termo RESISTÊNCIA À TRAÇÃO poderia ser subdividido em vinte termos mais específicos, cada um representando uma ordem de valores de resistência à tração) e eles estariam em algumas bases de dados, assim como os índices de uma empresa para seus próprios arquivos de contratos, índices de compilações de dados, ou certas bases de dados de patentes. Algumas das objeções de Breton, então, seriam contestadas mediante a indexação em nível muito mais alto de especificidade. Também é possível indexar as funções, desde que as que possivelmente se apliquem a um dispositivo sejam identificadas pelo autor e haja termos apropriados no vocabulário da base de dados. Porém, é totalmente irracional alimentar a expectativa de que o indexador seja capaz de reconhecer aplicações que não foram especificamente afirmadas pelo autor.

Posteriormente, Breton (1991) relatou pesquisas sobre um sistema de indexação que concretizava suas idéias e pretendia ajudar no processo de 'invenção'. O sistema experimental resultou da indexação de milhares de produtos industriais segundo as funções que desempenham e seus 'atributos distintivos'. Os atributos incluíam coisas como 'mais leve', 'mais barato', 'mais seguro' e 'mais forte'.

Alguns autores sugerem que é possível melhorar a recuperação em certos contextos por meio da indexação somente de determinadas características de um texto. Por exemplo, Oh (1998) sugere que, em psicologia, a indexação apenas de 'fatos empíricos' (nomes de variáveis, valores de correlação e informação sobre o nível de significância) melhoraria as condições de recuperação. Embora uma indexação altamente especializada como essa seja justificável em raras situações, é improvável que seja uma exigência da maioria e provavelmente será muito mais dispendiosa do que uma abordagem mais convencional.

Virou moda nos últimos anos considerar o problema da recuperação da informação como sendo fundamentalmente uma questão de comparar o 'estado anômalo de conhecimento' de um consulente com o estado de conhecimento mais 'coerente' dos autores (ver, por exemplo, Belkin et al., 1982), implicando isso que os problemas residem mais na saída do sistema (busca) do que na entrada. Há um certo equívoco nisso. Se aceitamos que a indexação é mais eficiente quando se orienta para as necessidades de determinado grupo de usuários, a função do indexador será prever os tipos de pedidos para os quais determinado documento será provavelmente uma resposta útil. Talvez isso ainda seja mais difícil do que prever quais os tipos de documentos que têm probabilidade de corresponder de modo útil a determinado pedido, o que constitui, em certo sentido, a função de quem faz a busca. Poder-se-ia argüir, então, que o estado 'anômalo' de conhecimento aplica-se mais ao lado de entrada do sistema de recuperação do que à sua saída. Olafsen e Vokac (1983) vêem essa particularidade com clareza:

> O indexador tem de fazer conjeturas sobre quais consultas serão formuladas pelo futuro usuário do sistema. Independentemente do grau de habilidade aplicada a esse exercício de adivinhação, ainda assim serão conjeturas, e o usuário recorre ao sistema

> levando sua própria questão concreta, e as associações que faz podem ser diferentes das do indexador (p. 294).

Estes autores também cometem um exagero de simplificação ao se referirem às questões trazidas pelo usuário como 'concretas', quando, de fato, muitas delas estarão longe disso. Apesar de tudo, talvez estejam certos ao sugerirem que os problemas de uma eficiente entrada de dados num sistema de recuperação superam os problemas concernentes à saída. Conforme Fairthorne (1958) salientou, há muitos anos: "A indexação é o problema fundamental bem como o obstáculo mais dispendioso da recuperação da informação."

Em algumas aplicações da indexação talvez seja possível ser bastante mais preciso no que se refere ao que deva ser considerado 'indexável'. Ao tratar da indexação de uma enciclopédia, Preschel (1981) oferece as seguintes diretrizes:

> Toda informação textual de natureza substantiva deve ser indexada. Define-se como 'substantiva' a informação que abranja de 8 a 10 linhas de texto ou que seja *singular* ou *notável* e que quase com certeza não ocorra em outro lugar da enciclopédia (p. 2).*

Em outras situações nem sempre é possível tanta precisão.

Com efeito, a questão sobre de que trata um item torna-se muito mais difícil quando se examina a indexação de obras de criação, como textos de ficção ou filmes de longa-metragem, ou imagens em geral. Nesses contextos, a atinência será vista em próximos capítulos.

Naturalmente, toda a questão da 'atinência' tornou-se muito mais complexa no atual ambiente de hipertexto/hipermídia. Quando um item pode ser vinculado [*linked*] a muitos outros, já não existe mais clareza sobre onde um começa e o outro acaba. Um documento trata apenas daquilo com que lida diretamente, ou trata também dos tópicos abordados nos itens a ele associados? Pouco se escreveu sobre a indexação de hipertextos de per si, embora nela se toque com certa extensão na literatura de hipertexto/hipermídia. Savoy (1995) e Salton et al. (1997) examinam possíveis métodos para o estabelecimento automático de vínculos [*links*] de hipertexto, o que pode ser considerado uma forma de indexação automática. Em capítulos posteriores trataremos desse tema.

Tradução

Tradução, a segunda etapa da indexação de assuntos, envolve a conversão da análise conceitual de um documento num determinado conjunto de termos de indexação. A esse respeito, faz-se uma distinção entre indexação por *extração* (indexação derivada) e indexação por *atribuição*. Na indexação por extração, palavras ou expressões que realmente ocorrem no documento são selecionadas para representar seu conteúdo temático. Por exemplo, o item da figura 3 poderia ser indexado com os seguintes termos:

* Esta citação de um texto inédito é reproduzida com autorização de Funk & Wagnalls.

OPINIÃO PÚBLICA	ISRAEL
PESQUISAS POR TELEFONE	EGITO
ESTADOS UNIDOS	AJUDA
ATITUDES	PAZ
ORIENTE MÉDIO	

todos os quais aparecem no título ou no resumo. Uma forma primitiva de indexação derivada, conhecida como *Uniterm*, empregava apenas termos formados por uma única palavra para representar o conteúdo temático. Se fosse estritamente observado, o sistema Uniterm acarretaria alguns resultados esquisitos, como a separação de Oriente Médio em ORIENTE e MÉDIO.

A *indexação por atribuição* envolve a atribuição de termos ao documento a partir de uma fonte que não é o próprio documento. Os termos podem ser extraídos da cabeça do indexador; por exemplo, ele decidiria que os termos AJUDA EXTERNA e RELAÇÕES EXTERIORES, que não aparecem explicitamente em nenhum dos resumos, seriam termos bons de usar no documento da figura 3.

Mais freqüentemente, a indexação por atribuição envolve o esforço de representar a substância da análise conceitual mediante o emprego de termos extraídos de alguma forma de vocabulário controlado.

Vocabulários controlados

Um vocabulário controlado é essencialmente uma lista de termos autorizados. Em geral, o indexador somente pode atribuir a um documento termos que constem da lista adotada pela instituição para a qual trabalha. Comumente, no entanto, o vocabulário controlado é mais do que uma mera lista. Inclui, em geral, uma forma de estrutura semântica. Essa estrutura destina-se, especialmente, a:

1. controlar sinônimos, optando por uma única forma padronizada, com remissivas de todas as outras;
2. diferençar homógrafos. Por exemplo, PERU (PAÍS) é um termo bastante diferente de PERU (AVE); e
3. reunir ou ligar termos cujos significados apresentem uma relação mais estreita entre si. Dois tipos de relações são identificados explicitamente: as hierárquicas e as não-hierárquicas (ou associativas). Por exemplo, o termo MULHERES OPERÁRIAS relaciona-se hierarquicamente com MULHERES (como uma espécie deste termo) e com DONAS DE CASA (também uma espécie do termo MULHERES), bem como está *associado* a outros termos, como EMPREGO ou FAMÍLIAS MONOPARENTAIS, que aparecem em hierarquias bem diferentes.

São três os tipos principais de vocabulários controlados: esquemas de classificação bibliográfica (como a *Classificação Decimal de Dewey*), listas de cabeçalhos de assuntos e tesauros. Todos procuram apresentar os termos tanto alfabética quanto 'sistematicamente'. Nas classificações, o arranjo alfabético é secundário, na forma de um índice que remete para o arranjo principal, que é hierárquico. No tesauro, o arranjo explícito dos termos é alfabético, mas existe uma

estrutura hierárquica implícita, incorporada à lista alfabética por meio de remissivas. A tradicional lista de cabeçalhos de assuntos é similar ao tesauro por ser de base alfabética, mas difere dele porque incorpora uma estrutura hierárquica imperfeita e por não distinguir claramente as relações hierárquicas das associativas. Os três tipos de vocabulário controlam sinônimos, distinguem homógrafos e agrupam termos afins, mas empregam métodos um tanto diferentes para alcançar estes objetivos.

Um estudo mais completo dessas questões encontra-se em Lancaster (1986).

Indexação como classificação

Na bibliografia de biblioteconomia e ciência da informação, faz-se, às vezes, uma distinção entre as três expressões *indexação de assuntos*, *catalogação de assuntos* e *classificação*. *Catalogação de assuntos* refere-se comumente à atribuição de cabeçalhos de assuntos para representar o conteúdo total de itens bibliográficos inteiros (livros, relatórios, periódicos, etc.) no catálogo das bibliotecas. *Indexação de assuntos* é expressão usada de modo mais impreciso; refere-se à representação do conteúdo temático de *partes* de itens bibliográficos inteiros, como é o caso do índice de final de livro. Assim, uma biblioteca pode 'catalogar' um livro sob o cabeçalho de assunto CÃES, para indicar seu conteúdo temático global; o conteúdo pormenorizado somente é revelado pelo *índice de assuntos* no final do livro. A distinção entre as expressões *catalogação de assuntos* e *indexação de assuntos*, uma delas referindo-se a itens bibliográficos inteiros e a outra a partes de itens, é artificial, enganosa e incongruente. O processo pelo qual o conteúdo temático de itens bibliográficos é representado em bases de dados publicadas — em formato impresso ou eletrônico — é quase invariavelmente chamado de *indexação de assuntos*, quer se estejam examinando itens total ou parcialmente. Assim, o *índice de assuntos*, por exemplo, do *Chemical Abstracts* remete a livros ou relatórios técnicos inteiros, bem como a partes de itens bibliográficos (capítulos de livros, trabalhos publicados em anais de eventos, artigos de periódicos). Por outro lado, as bibliotecas podem optar por representar em seus catálogos partes de livros (por exemplo, capítulos ou artigos); a isto se denomina comumente *catalogação analítica*. Quando aplicada ao conteúdo temático, esta atividade seria a catalogação analítica de assuntos.

A situação fica ainda mais confusa ao se examinar o termo *classificação*. Os bibliotecários costumam empregar esta palavra para designar a atribuição de números de classificação (extraídos de um esquema de classificação — por exemplo, o Decimal de Dewey (CDD), o Decimal Universal (CDU), o da Library of Congress (LC)) — a itens bibliográficos, especialmente com a finalidade de arrumá-los nas estantes das bibliotecas, em móveis de arquivo, etc. O catálogo de assuntos de uma biblioteca, porém, pode ser organizado alfabeticamente (*catálogo alfabético de assuntos* ou *catálogo dicionário*) ou organizado segundo a seqüência de um esquema de classificação (*catálogo sistemático*). Suponhamos que o bibliotecário tome um livro e decida que trata de 'aves'. Ele lhe atribui o cabeçalho de assunto AVES. Alternativamente, pode atribuir o número de classificação 598. Muitos se refeririam à primeira operação como *catalogação de assuntos* e à segunda como *classificação*, uma distinção totalmente absurda. A confusão é ainda maior quando se percebe que *indexação de assuntos* pode envolver o emprego de um esquema de classificação ou que um índice impresso de assuntos pode adotar a seqüência de um esquema de classificação.

Estas diferenças terminológicas são muito inexpressivas e só servem para confundir (ver Acton, 1986, para um exemplo típico). O fato é que a *classificação*, em sentido mais amplo, permeia todas as atividades pertinentes ao armazenamento e recuperação da informação. Parte dessa confusão terminológica se deve à incapacidade de distinguir entre as etapas de *análise conceitual* e de *tradução* na indexação.

Suponhamos que um especialista em informação tenha em mão um item bibliográfico e decida que ele trata do assunto 'robôs'. A atividade intelectual que tal decisão implica é a mesma, qualquer que seja o tipo de documento que tenha em mão — livro, parte de livro, periódico, artigo de periódico, anais de evento, trabalho apresentado em evento, seja o que for. O especialista *classificou* o item, isto é, colocou-o na classe conceitual de 'documentos que tratam de robôs'.

Como vimos antes, o processo de *tradução* envolve a representação da análise conceitual mediante um termo ou termos extraídos de um vocabulário. Um termo atribuído a um item constitui simplesmente um *rótulo* que identifica determinada classe de itens. Esse rótulo poderia ser o termo INTELIGÊNCIA ARTIFICIAL, extraído de um tesauro, de uma lista de cabeçalhos de assuntos ou do próprio documento, uma palavra equivalente de outra língua, ou um rótulo como 006.3 extraído de um esquema de classificação.

O processo que consiste em decidir do que trata um item e de atribuir-lhe um rótulo que represente esta decisão é conceitualmente o mesmo, quer o rótulo atribuído seja extraído de um esquema de classificação, de um tesauro ou de uma lista de cabeçalhos de assuntos, quer o item seja uma entidade bibliográfica completa ou parte dela, quer o rótulo seja subseqüentemente arquivado em ordem alfabética ou em outra seqüência (ou, com efeito, não arquivado de modo algum), quer o objeto do exercício seja organizar documentos em estantes ou registros em catálogos, índices impressos ou bases de dados eletrônicas.

No campo do armazenamento e recuperação da informação, a *classificação* de documentos refere-se à formação de classes de itens com base no conteúdo temático. Tesauros, cabeçalhos de assuntos e esquemas de classificação bibliográfica são essencialmente listas dos *rótulos* com os quais se identificam e, porventura, se organizam essas classes. O processo da busca de informação implica decidir quais classes consultar num índice impresso, catálogo em fichas ou base de dados eletrônica. A busca pode compreender o exame de uma única classe (por exemplo, tudo que apareça sob o cabeçalho ROBÔS) ou abranger combinações de várias classes (por exemplo, itens que apareçam sob ROBÔS e também

sob INTELIGÊNCIA ARTIFICIAL). Quantas combinações são possíveis ou com qual facilidade várias classes podem ser combinadas é algo que depende muito do formato da ferramenta que estiver sendo utilizada na busca, principalmente se for impressa ou em formato eletrônico.

Em suma, a *indexação de assuntos* é conceitualmente idêntica à *catalogação de assuntos*. A atividade que isso compreende é a *classificação de assuntos*, ou seja, formar classes de objetos com base em seu conteúdo temático. Neste texto, emprega-se *indexação de assuntos* ou mesmo *indexação*, por razões de comodidade, para designar todas as atividades de classificação de assuntos.

Especificidade do vocabulário

A figura 5 mostra uma análise conceitual feita para um artigo de periódico, bem como a tradução desta análise conceitual em três tipos diferentes de vocabulário. O artigo trata da utilização de robôs na indústria, especificamente, suas aplicações na fabricação e manuseio de materiais. Também examina o emprego de técnicas de inteligência artificial no projeto e operação de robôs, bem como os problemas específicos inerentes a fazer com que os robôs se movimentem adequadamente (isto é, problemas de locomoção).

Com relação a todos esses aspectos, a análise conceitual pode ser traduzida efetivamente para qualquer um dos vocabulários. Observe-se que as idéias transmitidas pela análise conceitual da figura 5 são abrangidas *coletivamente* pelos grupos de termos listados nos três vocabulários. Por exemplo, os três números de classificação da CDD, tomados em conjunto, abrangem o conteúdo temático desse artigo, de modo claro e completo, embora não haja uma relação unívoca entre os elementos individuais da análise conceitual e os termos da CDD. Embora edições anteriores da CDD não permitissem muita síntese das notações (isto é, a construção de números), edições posteriores permitem isso cada vez mais. Assim, 670.4272 (robôs em operações de fabricação) pode ser subdividido por 004-006. Uma vez que 006.3 representa inteligência artificial, os números podem ser combinados para formar o número altamente específico 670.427263.

A análise conceitual da figura 5 é abrangida de modo igual, completa e especificamente, em cada vocabulário, quando se consideram grupos inteiros de termos. No nível de um único termo, é claro, existem de fato diferenças. Se apenas um termo pudesse ser atribuído a esse artigo, a CDD seria melhor do que os outros vocabulários, pois é possível construir um único número de classificação que expresse o tópico principal desse artigo.

Este exemplo ilustra dois aspectos importantes. Primeiro, o tipo de vocabulário controlado (esquema de classificação, cabeçalhos de assuntos, tesauro) não é o fator mais importante a influir na etapa de tradução da indexação. Muito mais importantes são o alcance (abrangência) e a especificidade do vocabulário. Neste exercício de indexação, os três vocabulários podem abranger o assunto muito bem, embora de modo um tanto diferente. O segundo aspecto que o exemplo ilustra é que, embora a especificidade seja uma propriedade muito importante de um vocabulário controlado, pode ser obtida de diferentes formas em diferentes vocabulários. É importante considerar, em especial, as propriedades de *combinações* de termos de indexação mais do que as propriedades de termos isolados.

Análise conceitual	Classificação Decimal de Dewey	Library of Congress Subject Headings	INSPEC Thesaurus
Robôs industriais		ROBOTS, INDUSTRIAL	INDUSTRIAL ROBOTS
Inteligência artificial	670.427263 Inteligência artificial aplicada a robôs em operações de fabricação	ARTIFICIAL INTELLIGENCE	ARTIFICIAL INTELLIGENCE
Operações de fabricação		MANUFACTURING PROCESSES – AUTOMATION	MANUFACTURING PROCESSES
Manuseio de materiais	621.86 Equipamento de manuseio de materiais	MATERIALS HANDLING	MATERIALS HANDLING
Locomoção	531.112 Cinemática	ROBOTS – MOTION	KINEMATICS

FIGURA 5
Análise conceitual traduzida em três vocabulários controlados

Vejamos, por exemplo, um artigo sobre os serviços de saúde mental. O Vocabulário *A* contém o descritor específico SERVIÇOS DE SAÚDE MENTAL, enquanto o Vocabulário *B* possui o termo SERVIÇOS DE SAÚDE, mas não o termo mais específico. Porém, *B* também inclui o termo SAÚDE MENTAL, de modo que a idéia de 'serviços de saúde mental' é abrangida especificamente pela indexação sob SERVIÇOS DE SAÚDE e SAÚDE MENTAL. Sobre este tópico, portanto, o Vocabulário *B* é tão específico quanto *A*. Os vocabulários *C* e *D* são menos específicos: *C* contém o termo SAÚDE MENTAL, mas não possui termo algum para serviços de saúde, enquanto *D* traz SERVIÇOS DE SAÚDE, mas carece de um termo para saúde mental, de modo que nenhum dos dois apresenta a possibilidade de expressar especificamente a idéia de 'serviços de saúde mental'. No momento de realizar uma busca nos sistemas representados pelos diferentes vocabulários, seria possível obter resultados efetivos em *A* e *B*, mas seria impossível limitar a busca em *C* e *D* — ou seria recuperado tudo sobre saúde mental, ou tudo sobre serviços de saúde.

Este capítulo tratou dos princípios da indexação apenas teoricamente, pois não usou como modelo nenhum serviço de informação. É provável que grandes serviços de informação produzam suas próprias diretrizes de indexação, que merecem ser examinadas para se ver como as regras são aplicadas em determinado contexto. Um bom exemplo a estudar é o manual de indexação e resumos do Sistema Internacional de Informação Nuclear (Bürk et al., 1996).

CAPÍTULO 3

A prática da indexação

Ao indexador raramente é dado o luxo de poder ler um documento atentamente do começo ao fim. A exigência de indexar determinada quantidade de itens por dia haverá de lhe impor que se satisfaça comumente com uma leitura que estará longe de ser completa. Usualmente, recomenda-se um misto de ler e 'passar os olhos' pelo texto. As partes a serem lidas atentamente são as que apresentam maior probabilidade de dizer o máximo sobre o conteúdo no menor tempo: título, resumo, sinopse e conclusões. Os títulos das seções e as legendas das ilustrações ou tabelas também merecem maior atenção. Convém passar os olhos pelo restante do texto, para confirmar se as partes mais condensadas contêm uma imagem exata do que trata o documento. No entanto, o indexador deve, habitualmente, levar em conta o documento inteiro (partes lidas, partes que foram vistas de relance), e os termos atribuídos precisam refletir o todo. A exceção seria quando somente parte do documento (por exemplo, um documento longo com múltiplos assuntos) interessasse ao grupo de usuários a ser atendido.

Jones (1976), citando Anderson (1971), salienta que certas partes de um documento são particularmente gratificantes para o indexador: "Parágrafos de abertura (de capítulos ou seções) e frases de abertura e encerramento de parágrafos parecem ser especialmente ricos em palavras indexáveis." Isso confirma as conclusões de Baxendale (1958) em seu trabalho sobre o desenvolvimento de processos de indexação automática de documentos.

Uma norma internacional sobre indexação de assuntos (*Methods for examining documents*, 1985) oferece outras instruções sobre como analisar um documento:

> Muitas vezes é impraticável fazer uma leitura completa, que nem sempre é necessária, porém o indexador deve assegurar-se de que nenhuma informação útil lhe passou despercebida. As partes importantes do texto devem ser examinadas cuidadosamente, dando-se especial atenção às seguintes:
>
> a) título;
> b) resumo, se houver;
> c) sumário;
> d) introdução, as frases e parágrafos de abertura de capítulos, e as conclusões;
> e) ilustrações, gráficos, tabelas e respectivas legendas;
> f) palavras ou grupos de palavras que apareçam sublinhados ou grafados com tipos diferentes.

> Todos esses elementos devem ser examinados e avaliados pelo indexador durante a análise que faz do documento. Não é recomendável fazer a indexação a partir exclusivamente do título, e, se houver um resumo, não deve ser visto como um substituto satisfatório do exame do texto. Os títulos podem ser enganosos; tanto os títulos quanto os resumos podem ser inadequados; em muitos casos nenhum dos dois é uma fonte confiável do tipo de informação que o indexador requer (p. 2).

Em seu abrangente estudo sobre como os indexadores realmente executam suas atividades, Oliver et al. (1966) descobriram que a maioria adota, efetivamente, um método de ler/passar os olhos:

> O maior grupo de indexadores (cerca de 85% do total) afirmou que examinam rotineiramente o documento inteiro. Esses indexadores, porém, salientaram que certas partes do documento eram examinadas mais atentamente do que outras. Essas partes incluíam resumo, introdução, sinopse, conclusões, metodologia, resultados e tabelas e gráficos. Se uma ou mais de uma dessas seções 'condensadas' fosse considerada adequada pelo indexador, ele poderia examinar de relance ou simplesmente 'folhear' outras partes do documento. Os principais motivos para examinar o corpo do documento foram para constatar se alguma coisa passara despercebida, oferecer maior profundidade da indexação, e dirimir quaisquer dúvidas ou questões (p. 4-14).

Posteriormente, Chu e O'Brien (1993) observaram que indexadores novatos utilizavam bastante os resumos, quando existiam, para determinar o assunto dos artigos. Embora hajam observado mais de cem indexadores, o estudo abrangeu apenas três artigos.

Tudo isso se apóia no pressuposto de que é possível ler o documento a ser indexado. Conforme ressalta a ISO 5963 (*Methods for examining documents*, 1985), procedimentos diferentes se aplicarão a outros tipos de itens:

> Documentos não-impressos, como os meios audiovisuais, visuais e sonoros, inclusive objetos tridimensionais, exigem procedimentos diferentes. Nem sempre é possível, na prática, examinar um registro em sua inteireza (por exemplo, projetando um filme). A indexação, então, é comumente feita a partir de um título e/ou de uma sinopse, embora ao indexador deva ser dada a oportunidade de assistir ou ouvir o que se acha gravado, caso a descrição escrita seja inadequada ou pareça inexata (p. 2).

Um livro de Šauperl (2002) descreve como os catalogadores nas bibliotecas identificam o assunto de um livro e escolhem os cabeçalhos e números de classificação que lhe serão atribuídos. Baseia-se na observação minuciosa do trabalho de doze pessoas.

A indexação de fontes em formato eletrônico apresenta problemas especiais. Browne (2001), por exemplo, chamou atenção para os problemas relativos à indexação de sítios da Rede:

> A primeira etapa na indexação de um sítio da Rede é ter uma noção do volume e do tipo de material a ser indexado. No caso das provas tipográficas de um livro, é possível segurá-las com uma das mãos e folheá-las rapidamente. Na Rede isso é impossível, de modo que se é obrigado a examinar sistematicamente o sítio, anotando o

> tipo de informação, a quantidade de informações e a qualidade dos vínculos de navegação. Verifica-se o tamanho dos arquivos em megabytes. Solicita-se ao responsável pelo sítio [*webmaster*] que forneça o maior número possível de informações sobre os arquivos, inclusive quantos autores colaboraram com páginas. Quanto mais autores, mais variações serão previstas, e maior será a amostragem a ser feita (p. 32).

O motivo para se examinar o documento é, naturalmente, a decisão sobre o que incluir na indexação (nas palavras de Preschel (1972), isso constitui a identificação da 'matéria indexável').*

Conforme sugerido no capítulo 2, o indexador, para fazer isso com eficiência, precisa conhecer muito bem os interesses da comunidade servida pelo índice. Numa instituição específica, os indexadores podem ser orientados no sentido de procurar nos documentos certos elementos predefinidos; caso ocorram, *deverão* ser incluídos na indexação. Conforme o tipo de instituição, esses elementos importantes incluem: materiais de fabricação, temperaturas envolvidas, grupo etário envolvido, nível de escolaridade, etc. Em certos casos, os elementos mais importantes são pré-impressos no formulário de indexação, lembrando ao indexador que os termos apropriados devem ser usados, se se aplicarem a determinado documento. Por exemplo, a National Library of Medicine emprega 'etiquetas' [*checktags*] desse tipo para indicar grupos etários, gênero, tipos de animais utilizados em experiências, etc.

Essa etapa de 'análise conceitual' da indexação não deve ser influenciada pelas características do vocabulário a ser usado na etapa de tradução. Isto é, o indexador decide, primeiramente, quais os assuntos que precisam ser representados; só depois (a todo momento talvez) é que verificará se o vocabulário permite ou não representá-los adequadamente. Em outras palavras, o indexador não deve ignorar um assunto porque sabe ou desconfia que não pode ser expresso adequadamente. É possível que um exame mais meticuloso do vocabulário mostre que estava equivocado. Ademais, uma função importante do indexador é contribuir para o aperfeiçoamento do vocabulário controlado, comunicando suas deficiências aos responsáveis por sua manutenção. É improvável que isso ocorra se o indexador for estimulado a 'pensar' com os termos controlados. A propósito, discordo totalmente da ISO 5963, que afirma que "tanto a análise quanto a transcrição devem ser realizadas com o auxílio de instrumentos de indexação, como tesauros e esquemas de classificação". A transcrição, é claro, não se realiza sem essas ferramentas, mas a análise independe totalmente delas.

Um fator afim a lembrar é que a terminologia usada pelo autor pode não corresponder exatamente aos termos do vocabulário controlado. Mesmo que os termos empregados pelo autor coincidam com os termos controlados, a maneira como são utilizados pode ser diferente. Por exemplo, um autor emprega o termo 'epidemiologia' de forma muito imprecisa, mas o vocabulário define-o de modo

* Ver Milstead (1984) para outras considerações sobre como examinar um texto para identificar sua 'matéria indexável'.

mais preciso, e sua atribuição será errônea, apesar de ter sido usado pelo autor. São as idéias com que lida o autor, e não as palavras por ele empregadas, que devem ser indexadas.

Hjørland (2001) trata da seguinte forma a etapa da tradução na indexação:

> Uma decisão posterior refere-se a quais descritores do vocabulário controlado serão atribuídos ao documento. Tal decisão pode (e deve) ser vista da perspectiva inversa: sob quais descritores pareceria relevante para o usuário encontrar esse documento? (p. 777).

Embora concorde inteiramente com que a indexação esteja relacionada às necessidades de determinado grupo de usuários, acho que Hjørland pode estar confundindo as etapas de análise conceitual e tradução. É durante a primeira que se identificam as necessidades dos usuários. Isto é, o indexador decide quais aspectos do documento provavelmente interessarão aos usuários. Em seguida, o indexador seleciona os termos controlados que melhor representam esses aspectos.

Exaustividade da indexação

Os fatores que influem no desempenho de um sistema de recuperação da informação e que são diretamente atribuíveis à indexação podem ser assim categorizados:

1. Política de indexação
2. Exatidão da indexação
   - Análise conceitual
   - Tradução

As decisões quanto à política são tomadas pelos gestores do serviço de informação, estando, portanto, fora do controle do indexador individual; os fatores relativos à exatidão se estão sob o controle do indexador individual.

A principal decisão política diz respeito à *exaustividade* da indexação, a qual corresponde, grosso modo, ao número de termos atribuídos em média. O efeito da exaustividade foi anteriormente exemplificado na figura 3. A indexação exaustiva implica o emprego de termos em número suficiente para abranger o conteúdo temático do documento de modo bastante completo. A indexação seletiva, por outro lado, implica o emprego de uma quantidade muito menor de termos, a fim de abranger somente o conteúdo temático principal do documento. Quanto mais termos forem utilizados para indexar um documento mais acessível ele se tornará e, provavelmente, mais vezes será recuperado. Um centro de informação procurará indexar exaustivamente se seus usuários solicitarem com freqüência a realização de buscas completas. Um consulente que precise localizar todos os itens que, de alguma forma, tratem da OLP terá a expectativa de recuperar o documento mostrado na figura 3, mas isso somente será possível se a indexação tiver sido razoavelmente exaustiva.

As decisões da política, no que se refere à exaustividade, não devem assumir a forma de limites absolutos à quantidade de termos a serem atribuídos. Ao invés disso, essa política poderia sugerir uma faixa de termos; por exemplo, 'a maioria dos documentos será indexada com 8 a 15 termos'. Num grande centro de informação, que lide com muitos tipos diferentes de documentos, a política pode variar segundo o tipo de documento. Por exemplo, o centro de informação de uma grande empresa estabeleceria a seguinte política:

Relatórios técnicos da própria empresa	15–25 termos
Outros relatórios técnicos	10–15 termos
Patentes	15–20 termos
Artigos de periódicos	5–10 termos

e assim por diante. Alternativamente, a política tomaria como base o conteúdo temático, sendo os assuntos de maior interesse da empresa indexados com uma quantidade maior de termos.

Embora uma base de dados indexada exaustivamente costume possibilitar buscas exaustivas (alta *revocação*),* é provável que a indexação exaustiva saia mais cara do que a indexação seletiva. Ademais, a indexação exaustiva redundará em menor *precisão* das buscas. Quer dizer, será recuperado um número maior de itens que o consulente considera como não sendo pertinentes a sua necessidade de informação. Isso pode acontecer devido a dois motivos:

1. O número de 'falsas associações' aumentará conforme aumente o número de termos atribuídos. Por exemplo, o item da figura 3 seria recuperado durante uma busca sobre pesquisas por telefone no Egito, embora nada tenha a ver com este tópico.
2. Quanto mais termos forem empregados para indexar um item, mais ele será recuperado em resposta a assuntos de busca que nele são tratados somente de forma muito secundária. É provável que o item da figura 3 seja recuperado numa busca de artigos que tratem de líderes políticos dos estados árabes, porém a pessoa que solicita essa busca pode decidir que ele contribui tão pouco para este tema que dificilmente seria considerado útil.

A idéia de 'exaustividade' também se aplica a um sistema de recuperação que funcione com base em buscas feitas em textos (ver capítulo 13). O título do documento da figura 3 não constitui uma representação muito exaustiva de seu conteúdo temático. A exaustividade cresce à medida que aumenta o número de palavras presentes na representação.

O termo 'profundidade' é freqüentemente empregado para designar a quantidade de termos atribuídos a um documento. Quer dizer, emprega-se 'profundidade' em lugar de 'exaustividade'. Ambos os termos são imprecisos e podem ser enganosos. Para compreender melhor o efeito do aumento do número de termos

* Isso foi demonstrado em numerosas ocasiões; por exemplo, por Boyce e McLain (1989).

usados na indexação de um documento, imaginemo-lo como se possuísse duas dimensões, como mostra a figura 6. Digamos que o indexador consiga identificar dez assuntos afins que são estudados no documento. Considera-se isso como sendo o âmbito de abrangência do documento. Se o indexador tentar incluir todos esses assuntos, a indexação será tida como *exaustiva* (isto é, ela é uma representação exaustiva do conteúdo temático). Quanto mais assuntos forem incluídos mais exaustiva será a indexação. Por outro lado, quanto menos assuntos forem incluídos mais *seletiva* será a indexação. Evidentemente, a indexação exaustiva exigirá o emprego de maior número de termos.

A segunda dimensão do documento, do ponto de vista da indexação, é denominada *especificidade* na figura 6. Isto é, alguns assuntos identificados seriam indexados em mais de um nível de especificidade. Suponhamos que o primeiro assunto seja 'arquitetura de catedrais', que seria indexado sob o termo ARQUITETURA RELIGIOSA, que não é suficientemente específico. Para aumentar a especificidade, o indexador acrescentaria um segundo termo, CATEDRAIS. O emprego conjunto dos dois termos representa precisamente o assunto estudado. Por outro lado, a inclusão de ARQUITETURA DA HABITAÇÃO aumentaria a exaustividade e não a especificidade. pois estaria introduzindo um novo conceito na indexação.

Em outras palavras, a inclusão de mais termos de indexação aumentaria a exaustividade de uma representação ou aumentaria sua especificidade. Por conseguinte, embora seja verdadeiro dizer que a 'exaustividade' corresponde *grosso modo* ao número de termos atribuídos, não há uma relação unívoca exata entre exaustividade e número de termos. Neste livro, 'exaustividade' refere-se ao âmbito de abrangência da indexação exemplificado na figura 6. 'Profundidade' é um termo menos satisfatório porque denota o oposto de abrangência e se aplica de modo mais apropriado à dimensão da especificidade mostrada na figura 6.

A quantidade de termos atribuídos ao documento constitui realmente uma questão de custo–eficácia. Em geral, quanto mais exaustiva for a indexação maior será o custo,* e não é muito razoável indexar num nível de maior exaustividade que as necessidades dos usuários do serviço não justifique. Será preciso um nível mais alto de exaustividade se forem formulados muitos pedidos de buscas realmente exaustivas. No caso de serem feitos muitos pedidos de buscas que realmente cubram o assunto de modo completo, será necessário um alto nível de exaustividade. Se essas buscas que procuram exaurir o assunto forem a exceção e não a regra, bastará um nível muito mais baixo de exaustividade.

* Na realidade, naturalmente, isso é um exagero de simplificação. Quando tem em mãos um documento prolixo, o indexador talvez precise de mais tempo para incluir de modo exaustivo seu conteúdo. Em outros casos, talvez seja mais rápido usar muitos termos ao invés de tentar selecionar alguns poucos de um grupo em que eles podem estar estreitamente relacionados ou serem coincidentes. Em geral, no entanto, quanto mais termos forem usados, mais dispendioso será dar-lhes entrada na base de dados e processá-los subseqüentemente. Além disso, aumentar a quantidade de termos aumentará substancialmente os custos dos índices em formato de fichas ou impressos.

É claro que quanto mais termos forem empregados por documento (isto é, maior for a exaustividade), maior será a probabilidade de ele ser recuperado e maior será o número de características que o distingam de outros documentos. Mas a distribuição de itens entre os termos também afetará a discriminação: termos que se aplicam a muitos documentos não oferecerão muita discriminação; os que se aplicarem a poucos documentos serão bons discriminadores.

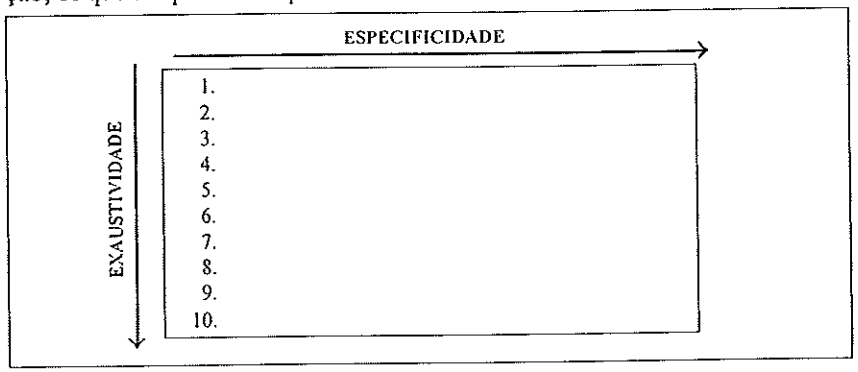


FIGURA 6
As duas dimensões da indexação de um documento

Wolfram e Zhang (2002) empregaram simulação em computador para estudar os efeitos da variação tanto dos níveis de exaustividade quanto das distribuições dos termos (o número médio de itens aos quais um termo se aplica). Sua conclusão foi que:

> Baixa exaustividade e distribuições de termos pouco profundas produzem menos diferenciação entre documentos, pois menos termos são atribuídos por documento e mais termos comuns são compartilhados entre documentos, disso resultando maior número de termos de baixo valor representativo. Exaustividade mais alta oferece maiores oportunidades para o acréscimo ao documento de termos adicionais mais distintivos. Igualmente, uma distribuição de termos mais profunda, quando se encontra um índice médio menor de atribuição de termos específicos ao item, acentua a distinguibilidade do documento. Quando se têm alta exaustividade para descrever de modo mais completo o documento e distribuições de termos mais profundas, definindo conjuntos de termos mais exclusivos, encontram-se as menores densidades de documentos, o que facilita distinguir os documentos entre si. Os resultados de cada modelo demonstram ser possível obter densidades espaciais similares de documentos com diferentes combinações de exaustividade de indexação e distribuições de termos. Por exemplo, a combinação de alta exaustividade/distribuição de termos pouco profunda e baixa exaustividade/distribuição de termos profunda resultou em níveis similares de densidade espacial de documentos (p. 950-951).

Os métodos de indexação e redação automáticas de resumos (ver capítulo 15) baseiam-se grandemente em critérios estatísticos (a contagem de ocorrências de palavras no texto), de modo que é possível aplicar medidas de 'densidade' (isto é, o número de termos de indexação ou a extensão do resumo em relação à extensão do texto). Connolly e Landeen (2001) propõem e aplicam medida similar (número de entradas do índice em relação ao número total de linhas de texto) aos índices do final de livros.

É óbvio que, à medida que as bases de dados crescem de tamanho, a quantidade de itens que aparecem sob qualquer termo também tende a crescer. Torna-se necessário, portanto, indexar com o emprego de mais termos (e também torná-los cada vez mais específicos) de modo que a indexação seja mais discriminativa para possibilitar pesquisas em que se alcance um nível adequado de revocação com nível tolerável de precisão. Lamentavelmente, isso não tem sido levado em conta na prática da catalogação de assuntos entre a comunidade bibliotecária dos EUA. O conteúdo temático dos livros é representado em nível muito genérico e superficial (em média, menos de duas combinações de cabeçalho de assunto/subcabeçalho por item, conforme O'Neill e Aluri, 1981). Mesmo que isso fosse aceitável há 50 anos, quando os acervos eram bem menores, e ainda o seja no caso de acervos muito pequenos, é hoje praticamente inútil em catálogos que abranjam vários milhões de itens. A conversão de catálogos em fichas para catálogos em linha proporcionou aos usuários uma grande vantagem potencial — a possibilidade de fazer buscas com termos em combinações lógicas. O valor potencial disso, porém, reduz-se grandemente devido ao baixo nível de exaustividade das representações constantes do catálogo. Por conseguinte, as pesquisas em linha nos catálogos de grandes bibliotecas universitárias freqüentemente resultam na recuperação de centenas de itens, a maioria dos quais talvez seja totalmente imprestável para o consulente (Lancaster et al., 1991). Esse 'fenômeno da recuperação volumosa' estimulou a realização de muitas experiências sobre como fazer buscas em grandes catálogos de forma mais discriminativa (ver, por exemplo, Prabha, 1991), tais como a delimitação por data, língua e outros critérios. O fato de a maioria dos catálogos permitir buscas nas palavras dos títulos (e às vezes nos números de classificação), bem como nos cabeçalhos de assuntos, parece ter tido, surpreendentemente, reduzido efeito na exaustividade da representação, uma vez que as palavras dos títulos, os cabeçalhos de assuntos e os números de classificação em geral se repetem (Xu e Lancaster, 1998).

Vários estudos examinaram a extensão com que os cabeçalhos de assuntos nos catálogos em linha de acesso público [OPACs] repetem as palavras-chave dos títulos dos livros. Voorbij (1998), por exemplo, analisou essa questão num contexto holandês. De fato, ele procurava comprovação de que a atribuição de descritores aos livros, um processo dispendioso, valia a pena. Ou seja, em que eles contribuem que as palavras-chave do título não o façam? Os descritores de assuntos conseguiram recuperar quase duas vezes mais itens relevantes do que as palavras-chave. Não só muitos títulos são indicadores inadequados daquilo de que trata um livro, mas, salienta Voorbij, o mesmo assunto pode aparecer nos

títulos representado de muitas maneiras diferentes. O controle de vocabulário imposto pelos cabeçalhos de assuntos é importante. Esse estudo foi realizado nas humanidades e ciências sociais, que podem, em média, apresentar títulos menos descritivos ou completos do que acontece nas ciências rígidas.

A figura 7 mostra a lei dos rendimentos decrescentes aplicada à indexação. No exemplo hipotético desse serviço de informação, a atribuição em média de X termos satisfará a cerca de 80% das necessidades dos usuários. A fim de elevar esse percentual para 90–95% seria preciso uma exaustividade muito maior na indexação. A posição do ponto X nessa curva e o que X representa em número de termos dependerão muitíssimo de questões específicas do sistema. Os gestores do serviço de informação elaboram diretrizes sobre exaustividade da indexação que resultam do seu conhecimento das necessidades dos usuários. Essas diretrizes costumam basear-se na intuição, embora seja possível realizar experimentos controlados em que se comparem amostras de necessidades de informação com uma coleção de documentos indexados com quantidades variadas de termos.

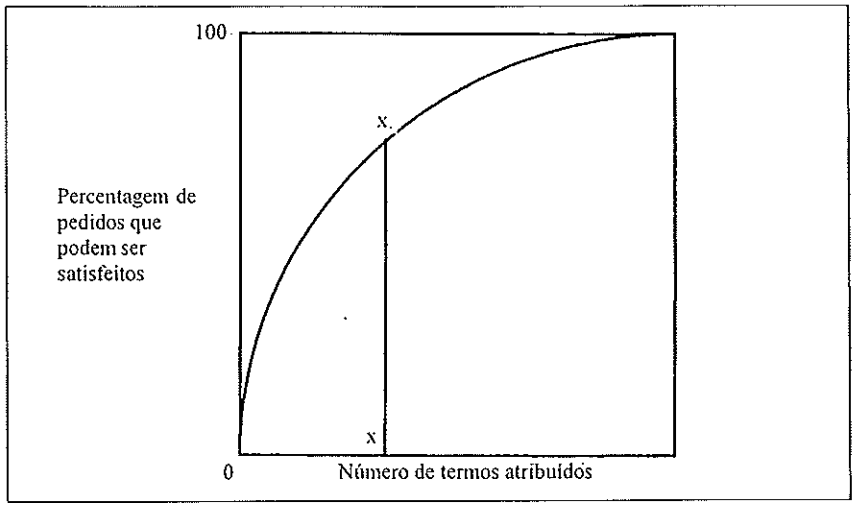


FIGURA 7
Rendimentos decrescentes na indexação

Evidentemente, a idéia de um nível ideal de exaustividade aplicável a todos os itens de uma base de dados é um tanto enganosa, uma vez que se aplicariam valores ideais extremamente diferentes a diferentes documentos, dependendo dos pedidos efetivamente formulados pelos usuários do sistema (Maron, 1979). A exaustividade ideal é inteiramente dependente dos pedidos.

Para alguns tipos de itens indexáveis, será excepcionalmente difícil chegar a um acordo sobre 'de que eles tratam'. Em relação a eles, não será fácil alcançar consenso e coerência na indexação, e, por isso, talvez precisem ser indexados em

nível exaustivo que atenda a diferentes pontos de vista. Isso acontece, por exemplo, na indexação de imagens, que podem ser vistas pelos indexadores em níveis diferentes, desde o altamente concreto até o altamente abstrato (Enser, 1995).

Intner (1984) mostrou que, ao decidir sobre o que incluir e quantos termos usar, o indexador poderia realmente exercer uma forma de censura, talvez deixando de cobrir algum aspecto do trabalho que ele desaprova. A situação inversa, naturalmente, consiste em usar uma quantidade injustificável de termos para fazer com que um item seja recuperado o maior número possível de vezes, tentação capaz de ocorrer se estiver associado à sua recuperação algum ganho financeiro ou de outro tipo. Tal fenômeno tem sido observado no ambiente da internet (ver capítulo 16), onde é denominado '*spoofing*' ou '*spamming*'. Price (1983) talvez tenha sido o primeiro a reconhecer isso como um problema em potencial.

Bell (1991a) estuda uma situação semelhante em relação aos índices do final de livros: os indexadores, ou possivelmente os editores, podem revelar um viés ao omitir certos tópicos do índice, ao reforçar as idéias do autor (ou o contrário), ou ao introduzir as próprias atitudes do indexador. Apresenta vários exemplos.

A quantidade de termos atribuídos a um documento é um fator crítico para definir se determinado item será ou não recuperado. Outros fatores afins, porém, também participam disso. Obviamente, é quase certo que a quantidade de itens recuperados venha a declinar à medida que mais termos forem combinados numa relação do tipo *e*, numa estratégia de busca. É claro que a extensão com que os termos podem ser combinados com êxito numa busca depende grandemente da quantidade de termos usados na indexação. Tomando-se um exemplo trivial, a combinação de três termos (A•B•C) pode recuperar um grande número de itens quando se emprega na indexação uma média de 20 termos por item, mas é improvável que recupere muitos deles numa base de dados em que somente três termos sejam atribuídos em média a cada item. (Devido a razões antes mencionadas, não recuperaria muitos, se é que recuperaria algum, numa busca feita num catálogo de biblioteca em linha.) Quanto mais seletiva for a indexação mais necessidade haverá de combinar termos numa relação do tipo *ou*, a fim de melhorar a revocação. As interações entre exaustividade de indexação e as características das estratégias de busca foram estudadas por Sparck Jones (1973). Estudos sobre o efeito da exaustividade em métodos mais automáticos de recuperação (ver capítulo 15) encontram-se em Shaw (1986, 1999a,b) e Burgin (1991, 1995).

Em muitos serviços de informação a indexação visa a duas finalidades um tanto diferentes: a) permitir que se tenha acesso a um item num índice impresso, e b) permitir que se tenha acesso a esse mesmo item numa base de dados eletrônica. Nessas circunstâncias, exige-se do indexador que indexe de acordo com certo nível de exaustividade preestabelecido para a segunda das finalidades acima, e que selecione um subconjunto dos termos de indexação (talvez entre dois e quatro) assim atribuídos, os quais servirão de pontos de acesso no índice impres-

so. Os termos desse subconjunto serão aqueles que o indexador considerar como os que melhor representam os aspectos mais importantes do documento. Isto pode ser visto como uma forma tosca de indexação 'ponderada': um termo recebe um de dois pesos — 'principal' (conteúdo temático fundamental, para o índice impresso) ou 'secundário' (todos os outros termos). No capítulo 11 examina-se mais detidamente a indexação ponderada.

Princípio da especificidade

O princípio que, isoladamente, é o mais importante da indexação de assuntos, e que remonta a Cutter (1876), é aquele segundo o qual um tópico deve ser indexado sob o termo mais específico que o abranja completamente. Assim, um artigo que trate do cultivo de laranjas será indexado sob LARANJAS e não sob FRUTAS CÍTRICAS ou FRUTAS.

Normalmente, seria melhor utilizar vários termos específicos, ao invés de um termo que seja mais genérico. Se um artigo descreve o cultivo de limões, limas e tangerinas, será mais bem indexado sob os três termos específicos do que sob o termo mais genérico FRUTAS CÍTRICAS. O termo FRUTAS CÍTRICAS será usado apenas para artigos que tratem das frutas cítricas em geral, e para aqueles que tratem praticamente de todas as frutas cítricas. Esta diretriz pode ser estendida à situação na qual se trata de várias frutas cítricas, mas não com muitos detalhes (a juízo do indexador) que justifiquem o emprego dos termos específicos. Em alguns casos, também, a clientela atendida pelo indexador pode estar interessada apenas em determinadas frutas. Nesta situação seria válido indexar apenas estas e não incluir termos correspondentes às outras frutas.

Alguns estudantes de indexação cometem o equívoco de indexar de modo redundante. Tendo indexado um artigo sobre laranjas sob o termo LARANJAS, sentem necessidade de também atribuir-lhe o termo FRUTAS CÍTRICAS e até mesmo FRUTAS. Não há necessidade disso. Na verdade, trata-se de uma prática de indexação medíocre. Se os termos genéricos forem atribuídos toda vez que for utilizado um termo específico, ficará difícil diferençar artigos genéricos de artigos específicos. Por exemplo, o usuário que consulta um índice sob o termo FRUTAS espera encontrar itens sobre frutas em geral, e não sobre frutas específicas.

Nos sistemas manuais de recuperação que antecederam os sistemas informatizados, de fato era preciso desdobrar as entradas dos termos específicos para os genéricos respectivos; por exemplo, o emprego do termo LARANJAS ao se indexar um item implicava que também lhe seriam atribuídos os termos FRUTAS CÍTRICAS, FRUTAS e talvez até mesmo PRODUTOS AGRÍCOLAS. A razão disso era permitir as buscas genéricas. Se não fosse assim, seria praticamente impossível realizar uma busca completa sobre, por exemplo, todas as frutas. Quando, no entanto, se projeta um sistema informatizado de modo apropriado, torna-se desnecessário esse desdobramento para os níveis genéricos, pelo menos quando se utiliza um vocabulário controlado. Convém, por exemplo, que haja a possibilidade de solicitar ao computador que faça uma busca sobre o termo FRUTAS e *tudo que estiver abaixo dele na estrutura hierárquica* (todos os termos específicos, TEs, no caso de um tesauro).

Em geral, portanto, não se deve contar com que os termos FRUTAS CÍTRICAS e LARANJAS sejam aplicados ao mesmo item. A única situação que justificaria esta combinação seria aquela onde houvesse um artigo que tratasse de frutas cítricas em geral, mas que incluísse extensas considerações sobre laranjas, ou outro que tratasse de frutas cítricas e em que as laranjas fossem o exemplo (por exemplo, a irrigação de frutas cítricas com exemplos tomados da irrigação de laranjais).

O indexador deve ter em mente que é possível conseguir especificidade mediante combinações de termos. Se não houver nenhum termo que sozinho possa representar o tópico, busca-se uma combinação apropriada de termos no vocabulário controlado. Eis alguns exemplos hipotéticos:

Literatura Francesa Medieval
    indexado sob LITERATURA MEDIEVAL e LITERATURA FRANCESA
Bibliotecas Médicas
    indexado sob BIBLIOTECAS ESPECIALIZADAS e CIÊNCIAS MÉDICAS
Literatura Canadense
    indexado sob LITERATURA e CANADÁ
Óleo de Amendoim
    indexado sob ÓLEOS VEGETAIS e AMENDOIM

Observe-se que o indexador deve procurar a combinação mais apropriada para cada caso. Teoricamente, Literatura Medieval Francesa seria expresso por meio de LITERATURA MEDIEVAL e FRANÇA, mas a combinação de LITERATURA MEDIEVAL e LITERATURA FRANCESA exprime a idéia de modo mais exato. Da mesma forma, combinou-se CIÊNCIAS MÉDICAS com BIBLIOTECAS ESPECIALIZADAS e não com BIBLIOTECAS, para expressar a idéia de bibliotecas médicas, pois estas são evidentemente especializadas, e combinou-se AMENDOIM com ÓLEOS VEGETAIS e não com ÓLEOS, uma vez que o óleo de amendoim é um óleo vegetal.

Às vezes, o vocabulário controlado não inclui um termo no nível de especificidade exigido por determinado documento. Nesse caso o indexador adotará o termo mais específico existente (por exemplo, FRUTAS CÍTRICAS, ao invés de FRUTAS, para um artigo sobre laranjas). Ele pode também sugerir à equipe responsável pela manutenção do tesauro que existe a necessidade de termos mais específicos nessa categoria.

Outras diretrizes

O processo da indexação de assuntos parece ser refratário a regras rigorosas. Além do princípio da especificidade, não foram desenvolvidas regras verdadeiras sobre a atribuição de termos, apesar de haver muitas acerca do que fazer com os termos de indexação depois de atribuídos (por exemplo, como estabelecer a seqüência em que são listados, a fim de formar cabeçalhos num índice impresso).

Muitas 'teorias' sobre indexação foram formuladas, algumas das quais passadas em revista por Borko (1977), porém, costumam não ser teorias de verdade, e oferecem pouca ajuda prática para o indexador.

Fugmann (1979, 1985) apresentou vários axiomas sobre 'indexação e provisão de informação', mas nem todos têm relação direta com a indexação como tal. O único princípio de indexação verdadeiro até agora formulado, denominado 'indexação compulsória', afirma que o indexador deve utilizar os termos mais apropriados de que disponha para descrever o conteúdo temático de um documento. Como isso significa, normalmente, os termos mais específicos, trata-se essencialmente de uma reiteração do princípio da especificidade. A maior parte dos axiomas de Fugmann corresponde realmente a fatores que influem no desempenho de sistemas de recuperação da informação e não a elementos de uma teoria da indexação, embora vários deles tenham implicações para a indexação. Por exemplo, o axioma da definibilidade tem relação com a capacidade de definir clara e inequivocamente uma necessidade de informação. Isso pode, evidentemente, ser estendido à capacidade de definir o conteúdo temático de documentos de modo claro e inequívoco. O axioma da previsibilidade diz que o êxito de uma busca num sistema de recuperação depende grandemente da previsibilidade com que é descrito o conteúdo temático, o que aponta para a importância da coerência na indexação. O axioma da fidelidade diz que outro fator que influi no desempenho é a capacidade de definir com rigor e exatidão o conteúdo temático (das necessidades de informação e, por extensão, dos documentos), que tem a ver mais com o vocabulário usado para indexar do que com a própria indexação.

Não consegui, de fato, encontrar uma teoria verdadeira, qualquer que fosse, aplicável ao processo de indexação, embora haja algumas (ver, por exemplo, Jonker (1964)) relativas às características dos termos de indexação. Ademais, creio ser possível identificar apenas duas regras básicas da indexação: uma, que se refere à etapa de análise conceitual, e a outra, à etapa de tradução, a saber:

1. Inclua todos os tópicos reconhecidamente de interesse para os usuários do serviço de informação, que sejam tratados substantivamente no documento.
2. Indexe cada um desses tópicos tão especificamente quanto o permita o vocabulário do sistema e o justifiquem as necessidades ou interesses dos usuários.

Estas regras estão, naturalmente, sujeitas a interpretação. Por exemplo, o que 'substantivamente' de fato significa? Uma orientação possível diria que o assunto *X* deve ser indexado quando se supõe que a maioria dos usuários que buscam informações sobre *X* considerariam esse item como sendo de interesse. É claro que 'substantivamente' não é uma propriedade que possa ser expressa ou medida com rigor. Se um dado assunto merece ou não ser indexado é algo que dependerá grandemente de três fatores: a) a quantidade de informações apresentadas sobre o assunto, b) o grau de interesse no assunto, e c) a quantidade de informações já existentes sobre o assunto: uma menção breve e isolada de um composto merece ser indexada se se sabe que esse composto é bastante recente; anos depois seria necessário um volume muito maior de informações para justificar sua inclusão.

A expressão 'necessidades ou interesses dos usuários', na segunda regra, implica que o princípio da especificidade pode e deve ser modificado quando se sabe que os usuários de um sistema ou ferramenta de informação, em certas circunstâncias, seriam mais bem servidos por meio da indexação de determinado tópico em nível mais genérico. Por exemplo, numa base de dados de medicina os artigos de veterinária aplicada a cães seriam indexados sob os nomes das respectivas raças caninas. Por outro lado, artigos sobre o uso de cães em experiências de laboratório seriam simplesmente indexados sob CÃES, mesmo quando a raça específica fosse mencionada.

Um corolário da primeira regra acima é que assuntos que não sejam examinados no documento não devem ser considerados pelo indexador. Embora isso paréça óbvio e banal, não é necessariamente assim. Alguns indexadores, principalmente os que se consideram 'especialistas' num assunto, podem sentir-se tentados a ver num documento coisas que jamais passaram pelas intenções do autor (por exemplo, aplicações de um dispositivo que extrapolam as alegadas no documento). Embora uma das funções importantes de certos especialistas em informação (como os que atuam na indústria) seja chamar a atenção dos usuários do serviço de informação para aplicações potenciais, isso, de fato, não constitui função do indexador. É muito melhor que se atenha ao texto e às afirmações do autor. O *ERIC processing manual* de 1980 traz bons conselhos a respeito disso:

> Indexe o documento que tem em mãos, não o documento que o autor *gostaria* de ter escrito ou *pretende* escrever no futuro. Não confunda suposições ou menções a implicações e possibilidades com o verdadeiro conteúdo (p. VII-13).

'Resultados não alegados pelo autor' não devem, é claro, ser confundidos com resultados negativos, pois estes comumente merecem ser indexados. Por exemplo, se um estudo mostra que certo material não se presta para ser utilizado em determinada aplicação, a aplicação mencionada deve ser definitivamente incluída na indexação, caso sejam contemplados outros critérios (por exemplo, o volume de informações fornecidas).

Em aplicações mais especializadas, os indexadores podem ser estimulados a buscar inferências. Por exemplo, Schroeder (1998), reportando-se a experiência no General Motors Media Archives, ressalta a importância de uma 'camada de inferência' na indexação de imagens. Por exemplo, uma fotografia de determinado veículo pode mostrá-lo atravessando um terreno acidentado, sendo então necessário identificar não somente o veículo mas também empregar termos que indiquem sua capacidade de desempenho em locais pedregosos.

Klement (2002) faz uma distinção entre indexação de 'sistema aberto' e indexação de 'sistema fechado'. A última (cujo exemplo mais evidente são os índices de final de livro) refere-se a índices de um único item; esses índices são não-contínuos. A indexação de sistema aberto, ao contrário, aplica-se a inúme-

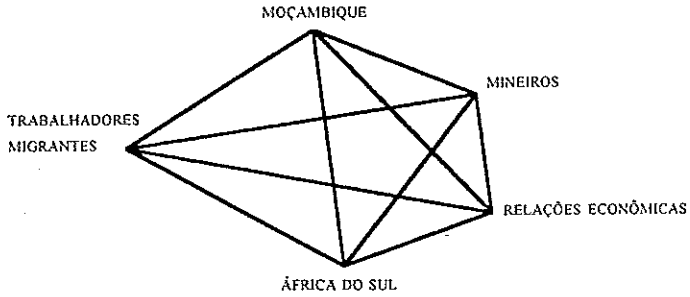
ros itens e é contínua, como é o caso da indexação de artigos de periódicos em bases de dados como, por exemplo, o MEDLINE. Quando a indexação se aplica a muitos itens, e é contínua, os termos adotados nas entradas do índice devem ser padronizados. A padronização não constitui de fato um problema na indexação de sistema fechado, embora seja obviamente necessário utilizar uma terminologia uniforme, coerente, em toda a extensão do índice. A indexação de sistema fechado pode usar termos que são não-contínuos: 'Leonardo da Vinci, morre' pode ser perfeitamente apropriado nesse tipo de índice, sendo improvável que apareça num índice de sistema aberto (embora 'Leonardo da Vinci' apareça).

Índices pós-coordenados

O conteúdo temático objeto de um documento e representado pelos termos de indexação que lhe são atribuídos possui caráter multidimensional. Vejamos, por exemplo, um artigo que trate da migração de mão-de-obra de Moçambique para as minas da África do Sul e que é indexado sob os seguintes termos:

MOÇAMBIQUE
ÁFRICA DO SUL
TRABALHADORES MIGRANTES
MINEIROS
RELAÇÕES ECONÔMICAS

Embora os termos sejam aqui apresentados em forma de lista, representam, na realidade, uma rede de relações:



Convém recuperar esse documento durante uma busca que envolva qualquer um dos termos tomados isoladamente ou qualquer combinação entre eles: quaisquer dois termos, quaisquer três, quaisquer quatro, ou todos os cinco. Um sistema de recuperação da informação que permite que uma busca combine os termos de qualquer maneira é freqüentemente denominado *pós-coordenado* (outras denominações empregadas têm sido *pós-combinação* ou *manipulatório*).

Os sistemas pós-coordenados surgiram na década de 1940, quando foram implantados com a utilização de vários tipos de fichas. Um sistema informatizado moderno, funcionando em linha, pode ser visto como um descendente direto

desses sistemas manuais. Pode-se imaginá-lo conceitualmente como uma matriz semelhante à mostrada na figura 8.

Os arquivos de um sistema em linha incluem dois elementos principais:

1. Um conjunto completo de representações de documentos: a referência bibliográfica acompanhada normalmente de termos de indexação ou um resumo, ou ambos.
2. Uma lista de termos que mostra quais os documentos indexados sob eles (às vezes chamada *arquivo invertido* ou *arquivo de lançamentos*). Os documentos são identificados por números de registro como mostra a figura 8.

Pode-se demonstrar o que se passa durante uma busca em linha consultando a matriz da figura 8. Suponhamos que quem faz a busca entra com MOÇAMBIQUE num terminal e que este termo é representado por P no diagrama. O sistema responde indicando que sete itens foram indexados sob tal termo. A pessoa entra com TRABALHADORES MIGRANTES (L no diagrama) e recebe a informação de que quatro itens aparecem sob este termo. Se ela pedir agora que seja feita a combinação de L com P, o sistema comparará os números dos documentos nas duas listas e indicará que três itens satisfazem a esse requisito. Atendendo à solicitação do interessado, o computador localiza esses registros pelos seus números de identificação (4, 8, 10) e os mostra na tela do monitor ou os imprime.

Esse processo permanece o mesmo independentemente de quantos termos se achem envolvidos e quais sejam as relações lógicas especificadas por quem faz a busca. Se for pedido F *ou* G, o sistema indicará que cinco itens satisfazem à condição. Quem faz a busca solicita então que esta lista de cinco itens seja combinada com a lista sob N — isto é, (F *ou* G) *e* N — do que resulta a recuperação de três itens. A respeito dos sistemas pós-coordenados é possível afirmar que:

1. Os termos podem ser combinados entre si de qualquer forma no momento em que se faz a busca.
2. Preserva-se a multidimensionalidade das relações entre os termos.
3. Todo termo atribuído a um documento tem peso igual — nenhum é mais importante do que outro (embora a indexação ponderada, estudada em capítulo posterior, possa ser utilizada).

Estas características não se aplicam a índices pré-coordenados, que serão objeto do próximo capítulo.

Instrumentos auxiliares da indexação

O indexador precisa contar com alguma forma de anotação dos resultados da operação de indexação. São quatro as possibilidades:

1. Anotação no próprio documento
2. Preenchimento de algum tipo de formulário impresso em papel
3. Gravação numa fita de áudio
4. Preenchimento de um formulário mostrado na tela de um monitor em linha

TERMOS (CLASSES)	DOCUMENTOS														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	X									X					
B		X			X		X		X						X
C					X	X					X				
D						X				X					
E	X		X				X	X	X					X	X
F	X														
G			X	X			X					X			
H	X										X			X	
I													X		
J	X		X				X		X					X	X
K	X	X	X				X			X	X	X			
L				X				X		X			X		
M		X				X				X					X
N	X			X	X			X	X			X	X		
O		X	X			X	X								X
P	X			X	X			X	X	X				X	

FIGURA 8
Sistema de recuperação da informação representado como uma matriz

Embora hoje em dia a norma seja indexar diretamente em linha, com o emprego de algum tipo de tela estruturada, antigamente eram comuns outras possibilidades que ainda são adotadas em certos lugares.

Em certas instituições o indexador simplesmente marca o documento que tem em mãos, e um datilógrafo transcreve o que ele marcou. Esta forma de trabalho só convém, normalmente, se se adota um método de indexação relativamente simples, como, por exemplo, enriquecimento de títulos associado à inclusão de uma quantidade relativamente pequena de termos ou códigos de indexação.

Até que os sistemas em linha se tornassem comuns, era corriqueiro o indexador dar entrada aos termos num formulário impresso. A figura 9, por exemplo, mostra uma versão do formulário que era adotado pela National Library of Medicine. Observe-se o emprego de 'etiquetas' [*checktags*], que são termos potencialmente aplicáveis a inúmeros documentos da base de dados. Sua pré-

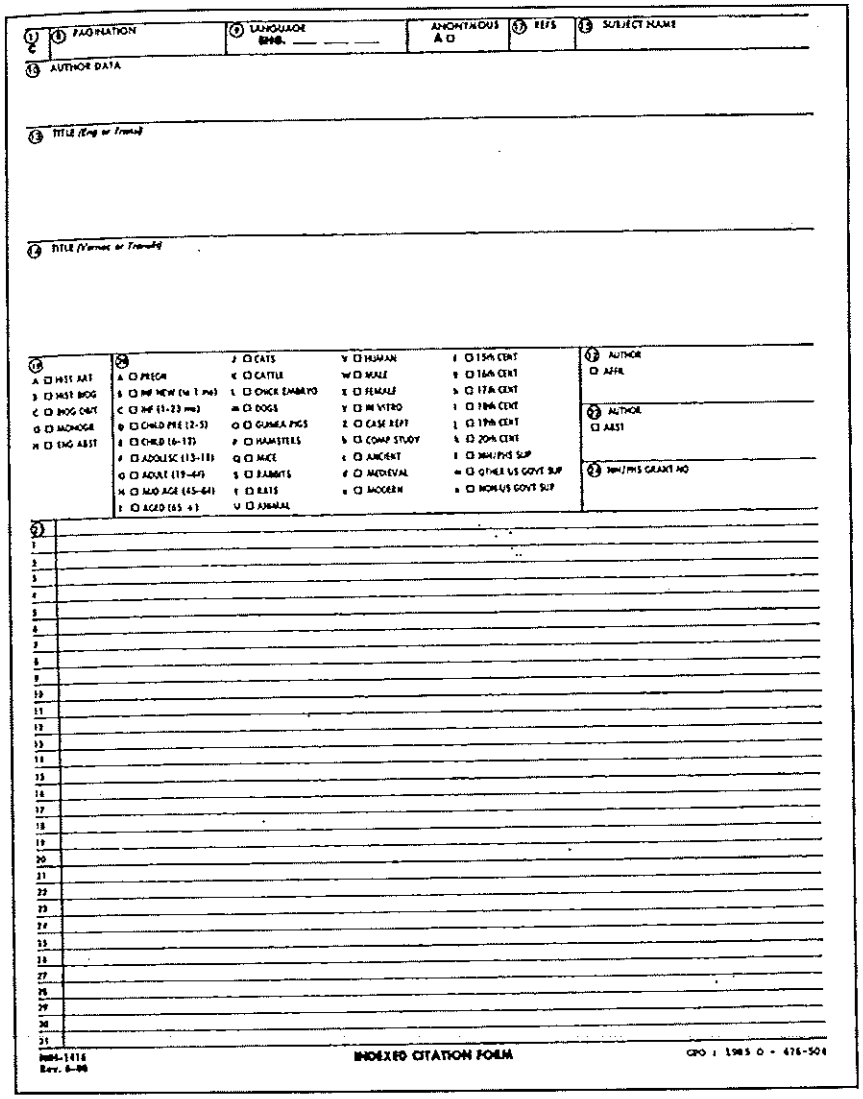


FIGURA 9
Formulário de indexação utilizado antigamente pela National Library of Medicine

impressão no formulário é eficiente e econômica, pois o indexador só precisa ticar os que se aplicarem a cada caso. Isto não só poupa tempo ao indexador mas também faz com que se lembre de que esses termos devem ser atribuídos sempre que se aplicarem a determinado documento. Devido a essa forma de advertência, as etiquetas são atribuídas de modo mais coerente do que outros termos (Lancaster, 1968a; Funk et al., 1983).

Qual o material estudado?	O processo é dinâmico (ao invés de estático)?	Existem cargas aerodinâmicas específicas?	Envolve resistência estrutural e elasticidade?
Metais	Vibrações	Sustentação	Tensão e deformação
Gases	Resposta transitória	Arrasto	Plasticidade
Plásticos	Impacto	Momento	Falha
Alumínio	Estabilidade	Rajada	Propriedades-limite
Magnésio	Velocidade	Pressão	Propriedades materiais
Titânio		Centro de aplicação, p. ex., centro aerodinâmico, centro de pressão, etc.	Aeroelasticidade
Ar			Vibração
Qual o tipo de escoamento de fluido?	**É um problema de estabilidade e controle?**	**Ou existe outro problema de aerodinâmica?**	**Existe um processo térmico envolvido?**
Escoamento de fluido	Estabilidade	Camada-limite	Termodinâmica
Escoamento interno	Controle	Aeroelasticidade	Constantes termodinâmicas
Subsônico	Estático	Vibração	Combustão
Transônico	Resp. trans. dinâmica	Deflexão	Transferência de calor
Supersônico	Longitudinal	Perda de sustentação	Resfriamento
Hipersônico	Lateral	Interferência	Convecção
Laminar	Derivadas	Hidráulica	Condução
Turbulência	Amortecimento	Trajetória	Térmico
Escoamento de escorregamento	Peso e equilíbrio, p. ex., centro de gravidade, momentos de inércia, etc.	Gotículas	Radiação
Compressibilidade		Técnica modificante	Aquecimento aerodinâmico
Viscosidade		Desempenho	
Vórtices			
Ondas de choque			
Envergadura finita			

FIGURA 10
Formulário característico da indexação de Mooers
Reproduzido de Brenner & Mooers (1958) com permissão de Van Nostrand Reinhold

Em ambientes onde se pratica uma indexação altamente especializada, talvez seja possível pré-imprimir o vocabulário controlado completo no formulário de indexação, permitindo assim que todos os termos se tornem basicamente etiquetas. O pioneiro desse método foi provavelmente Mooers. A figura 10 (conforme Brenner e Mooers, 1958) mostra um formulário característico da indexação de Mooers. Observe-se como os descritores são agrupados sistematicamente. Ao analisar o documento, o indexador considera basicamente cada descritor da tabela como potencialmente aplicável. Com efeito, o indexador formula a si mesmo as perguntas propostas pelo próprio formulário de indexação. Se, por exemplo, a resposta à pergunta 'existem cargas aerodinâmicas?' for 'sim' (isto é, o documento em exame trata de cargas específicas), o indexador levará isso em consideração atribuindo o descritor, ou descritores, mais apropriado para carga

aerodinâmica. A lista de descritores, apresentada dessa forma, simplifica o processo de indexação porque poupa ao indexador uma parte de seu esforço intelectual. As utilizações potenciais que um documento de interesse pode ter para a instituição são representadas pela lista de perguntas 'orientadoras' que foi criteriosamente compilada por pessoal científico graduado. O indexador simplesmente segue as 'dicas' dadas nessa lista.

Antigamente, o U.S. Patent and Trademark Office desenvolveu pequenos sistemas de recuperação limitados a uma única classe ou a um número restrito de classes na área de patentes. Foram criados vocabulários especializados para essas classes, suficientemente sucintos para que fossem impressos em poucas folhas. A figura 11 mostra parte de um desses vocabulários, destinado à subclasse de patentes que tratam de computadores digitais de uso geral. Igual às tabelas de descritores de Mooers, todo o vocabulário pode ser facilmente examinado, evitando que o indexador deixe passar despercebido um termo importante, e eliminando a necessidade de dar entrada aos termos num formulário de indexação. Neste caso, encontram-se disponíveis múltiplos exemplares da lista de termos, e uma patente é indexada simplesmente traçando-se um círculo em volta dos termos apropriados ou seus códigos num exemplar da lista. Todo o processamento posterior requer apenas trabalho de rotina. O 'microtesauro' do Air Pollution Technical Information Center, descrito por Tancredi e Nichols (1968), foi também criado para ser utilizado traçando-se um círculo em volta dos termos. Uma parte desse microtesauro é mostrada na figura 12.

Também se logrou êxito em algumas instituições onde o indexador passou a ditar os termos num gravador de fita para serem posteriormente transcritos por datilógrafos. Este método de fato apresenta alguns problemas. Podem ocorrer muitos erros de datilografia quando se emprega um extenso vocabulário técnico, estranho ao datilógrafo, o que exige um trabalho de revisão muito cuidadoso. Alguns indexadores não conseguem trabalhar bem dessa maneira porque têm dificuldades em se lembrar quais foram os termos que já atribuíram a um item.

Hoje em dia, porém, a maioria dos produtores de bases de dados adota processos de indexação em linha. Assim, aparecem no monitor várias telas formatadas e o indexador vai inserindo os dados nos campos apresentados. Essa modalidade de operação oferece grandes vantagens em relação às suas predecessoras: o indexador pode receber vários tipos de mensagens, alguns de seus equívocos podem ser reconhecidos por programas de detecção de erros que o advertem imediatamente, além de dispensar a etapa rotineira intermediária, quando se converte o trabalho do indexador para formato eletrônico. Ademais, existe a possibilidade de o indexador passar do modo de entrada de dados para o de recuperação, e assim valer-se de casos precedentes para se orientar quanto a certas decisões concernentes à indexação. Quer dizer, o indexador acessa a base de dados, para verificar como um termo foi usado antes ou como um documento mais antigo, afim a outro que está sendo examinado, foi indexado.

```
          SYSTEM ARCHITECTURE
228       .Plural processors with different
             internal structures (28/0)
228.1     .Shared memory (28/1)
228.2     .Virtual processor/machine (28/2)
228.3     .Plural (redundant) central processors
             (28/3)
228.4     .Central processor combined with
             terminal processor (28/4)
228.5     .Central processor combined with
             interface processor (28/5)
228.6     .Central processor combined with
             coprocessor (28/6) *
228.7     .Multiple instruction multiple data
             (MIMD) (28/7) *
228.8     ..Loosely coupled MIMD (28/8) *
228.9     ..Tightly coupled MIMD (28/9) *
229       .Multiprocessor interconnection (29/0)
229.1     ..Direct (29/1)
229.2     ..Parallel (common bus) (29/2)
229.3     ..Loop (29/3)
229.4     ..Reconfigurable (29/4)
229.41    ..Tree structure (29/A) *
229.5     ..Other specific multiprocessor
             interconnection (29/5)
230       .Multiprocessor/Processor control (30/0)
230.1     ..Priority assignment (30/1)
230.2     ..Interrupt handling (30/2)
230.3     ..Task assignment (30/3)
230.4     ..Supervisory (master/slave) (30/4)
230.5     ..Other specific multiprocessor control
             (30/5)
230.6     .Other specific multiprocessor system
             (30/6)
231       .Mini/Micro/Personal computer (31/0)
231.1     ..Portable (31/1)
231.2     ...Hand-held/Carried on person (31/2)
231.3     ...Other portable computer (31/3)
231.31    ..Other specific mini/micro/personal
             computer (31/A) *
231.4     .Timeshared (31/4)
231.5     ..Peripheral devices (31/5)
231.6     ..Plural programs (Multiprogrammed)
             (31/6)
231.7     ..Other specific timeshare (31/7)
231.8     .Pipelined (31/8)
231.9     .Parallel array/Single Instruction
             Multiple Data (SIMD) (31/9)
232       .Orthogonal (32/0)
232.1     .Virtual (32/1)
232.2     .Adaptive (32/2)
232.21    .Vector processor (32/A) *
232.22    .Data flow (32/B) *
```

FIGURA 11
Parte de vocabulário especializado sobre computadores digitais utilizado pelo U.S. Patent and Trademark Office
Reproduzida com permissão do U.S. Patent and Trademark Office

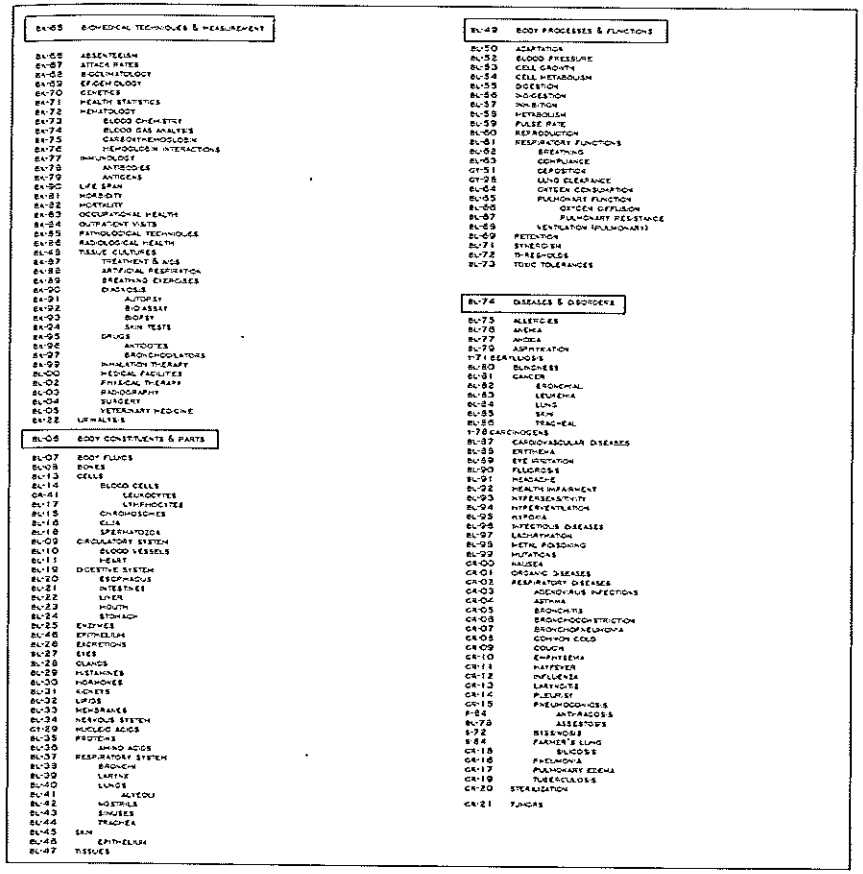


FIGURA 12
Seção do microtesauro do Air Pollution Technical Information Center
Apud *American Documentation* (Tancredi & Nichols [1968])

Um típico sistema de indexação em linha, conhecido como DCMS (Data Creation and Maintenance System), é utilizado pela National Library of Medicine para entrada de dados na base MEDLINE. O trabalho do indexador consiste em preencher várias 'telas' no monitor. Ver, por exemplo, a figura 13, que mostra uma tela com a versão atual das etiquetas. Observe-se que o indexador ticou (✓) as etiquetas que se aplicam a esse artigo do *American Journal of Human Genetics*, a saber, *adult*, *middle age*, *aged*, *human*, *male*, e *female*. A figura 14 mostra a tela seguinte com as etiquetas selecionadas pelo indexador. Vários descritores (cabeçalhos de assuntos sozinhos ou com subcabeçalhos) foram selecionados pelo indexador. O sistema oferece a possibilidade de enviar mensagens ao indexador. Por exemplo, se for usada a etiqueta *pregnancy*, o DCMS informará auto-

maticamente ao indexador para acrescentar *female* e o advertirá para usar *animal* ou *human*. O DCMS também advertirá para o emprego de certas etiquetas, com base num número limitado de palavras que ocorrem nos títulos ou resumos. Por exemplo, se a palavra '*feline*' aparecer no texto, o indexador será advertido para examinar a possibilidade de usar a etiqueta *cats*.

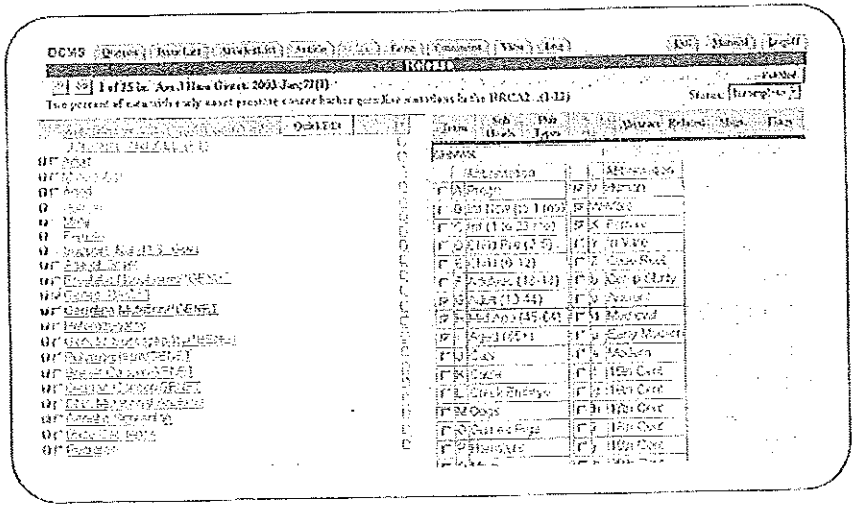


FIGURA 13
Tela de etiquetas no DCMS

O DCMS tem outras características que facilitam o processo de indexação. O vocabulário (*Medical Subject Headings*) pode ser visualizado na tela e o indexador pode selecionar termos sem ter que redigitá-los. Para qualquer um dos termos que for selecionado o sistema pode ser solicitado a mostrar na tela do monitor uma anotação explicativa ou, alternativamente, uma lista dos subcabeçalhos que podem ser usados com esse termo. O sistema também levará ('mapeará') de um termo não-aprovado para um aprovado por meio das remissivas incluídas no *Medical Subject Headings*.

Obviamente, o vocabulário controlado usado por um serviço de informação será ferramenta de importância crucial para o indexador. Deverá ser organizado e apresentado de forma a oferecer ao indexador um auxílio positivo na seleção dos termos mais apropriados para determinada situação. Embora tenha estreita relação com o tema da indexação, a construção e as propriedades dos vocabulários controlados são questões que estão fora do âmbito deste livro. Foram tratadas com detalhes em outras publicações (Lancaster, 1986; Soergel, 1974).

Um tesauro publicado incorpora normalmente um limitado *vocabulário de entradas* na forma de remissivas do tipo *ver*, *usar* ou *ver sob*. Um grande centro de informação poderá desenvolver um vocabulário de entradas separado para

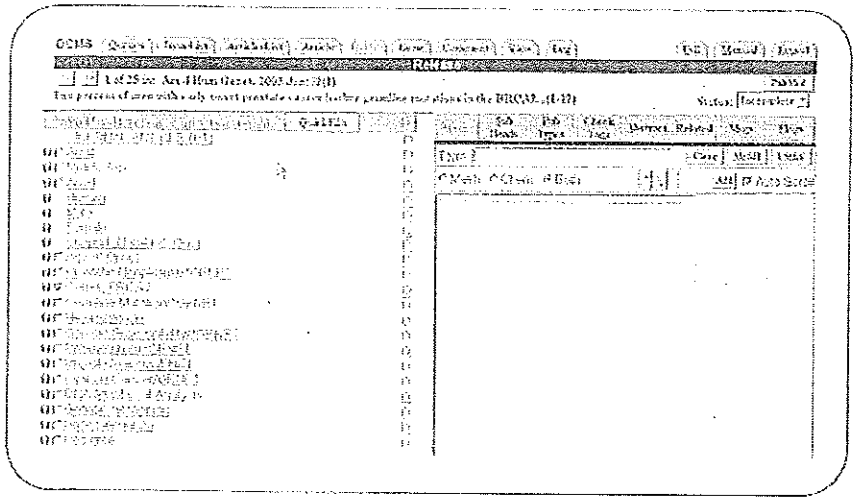


FIGURA 14
Registro de indexação pronto no DCMS

uso local pelos indexadores, consulentes e lexicógrafos. Esse vocabulário poderá estar disponível em formato impresso ou em linha.

Por exemplo, a National Library of Medicine (NLM) lança mão de várias ferramentas, ricas em componentes de vocabulário de entradas e diretrizes para indexação. A mais óbvia é o navegador eletrônico MeSH Browser. Esta ferramenta, feita para a Rede, destina-se a ser usada por indexadores, catalogadores de assuntos e especialistas em buscas e é muito mais elaborada do que os *Medical Subject Headings*, que tem por finalidade servir de guia no uso do *Index Medicus* impresso. Outra ferramenta, em formato impresso, é *Medical Subject Headings – Annotated Alphabetic List*. A figura 15 mostra algumas entradas desta versão anotada. Essa ferramenta bastante complexa possui componentes de vocabulário de entradas (por exemplo, *depth intoxication* [intoxicação profunda] remete para *inert gas narcosis* [narcose por gás inerte] bem como outras diretrizes ou instruções para indexação: termos relacionados (ver, por exemplo, o fato de que *depressive disorder* [transtorno depressivo] deve ser diferençado de *depression* [depressão], termos que eram usados antigamente (por exemplo, no período de 1973–1990 o termo *dermacentor* somente existia para buscas em linha; para impressão no *Index Medicus* esse organismo devia ser indexado também sob o cabeçalho mais genérico *ticks* [carrapatos]) e mesmo algumas definições (ver, por exemplo, *dermabrasion* [dermabrasão]).

Entre os vocabulários de entradas mais específicos desenvolvidos pela NLM está o *Tumor key* [Chave de tumores], que orientava sobre indexação de doenças neoplásicas. A figura 16 mostra entradas desse vocabulário. Observe-se como

Depressive Disorder
F3.600.300+
do not confuse with DEPRESSION: see note there; depression lasting over 2 years = DYSTHYMIC DISORDER
81; DEPRESSION, NEUROTIC was DEPRESSIVENEUROSES see DEPRESSION, REACTIVE 1979-80, see under DEPRESSION, REACTIVE 1969-78
X Depression, Endogenous
X Depression, Neurotic
X Depression, Unipolar
X Depressive Syndrome
X Melancholia
X Neurosis, Depressive
X Unipolar Depression

Depressive Disorder, Major see Depression, Involutional

Depressive Symptoms see Depression

Depressive Syndrome see Depressive Disorder

Depth Intoxication see Inert Gas Narcosis

Depth Perception
F2.463.593.200+ F2.463.593.932.869.255+
G11.697.911.860.317+
disord of depth perception: coord IM with PERCEPTUAL DISORDERS (IM)
X Stereopsis
X Stereoscopic Vision

Dequalinium
D3.438.810.824.200
1991(1976); see QUINOLINIUM COMPUNDS 1976-1990; for DECHALINIUM & DEQUALONUM see DEQUALINIUM 1976-1993

Dercum's Disease see Adiposis Dolorosa

Derealization see Depersonalization

Dermabrasion
E4.680.250
mechanical planing of the skin; do not use /util except MeSH definition

Dermacentor
B1.131.166.132.832.400.200
infestation: coord IM with TICK INFESTATIONS (IM)
91(73); was see under TICKS 1973-90

FIGURA 15
Exemplo de entradas de *Medical subject headings – annotated alphabetic list (2003)*

pode ser considerado um verdadeiro vocabulário de entradas que inclui tanto remissivas unidirecionais quanto multidirecionais. Por exemplo, cisto teratóide [*teratoid cyst*] deve ser indexado sob teratoma, porém cistoadenocarcinoma do ducto biliar [*bile duct cystadenocarcinoma*] será indexado sob *cystadenocarcinoma* [cistoadenocarcinoma] e também sob *cholangiocarcinoma* [colangiocarcinoma]. Esses vocabulários especializados não são mais mantidos pela NLM.

cyst, teratoid - TERATOMA

cyst, teratomatous - TERATOMA

cyst, thyroglossal
- THYROGLOSSAL CYST not neoplastic

cyst, umbilical
- URACHAL CYST (not neoplastic)

cyst, urachal - URACHAL CYST (not neoplastic)

cystadenocarcinoma (unspecified)
- CYSTADENOCARCINOMA

cystadenocarcinoma, bile duct
- CYSTADENOCARCINOMA + CHOLANGIOCARCINOMA

cystadenocarcinoma, endometrioid
- CARCINOMA, ENDOMETRIOID

cystadenocarcinoma, mucinous
- CYSTADENOCARCINOMA, MUCINOUS

cystadenocarcinoma, mucinous papillary
- CYSTADENOCARCINOMA, MUCINOUS

cystadenocarcinoma, papillary (unspecified)
- CYSTADENOCARCINOMA, PAPILLARY

FIGURA 16
Exemplo de entradas de *Tumor key*, um vocabulário de entradas especializado antigamente utilizado pela National Library of Medicine

A maioria dos tesauros publicados inclui componentes de vocabulários de entradas, mas é improvável que possuam a riqueza (ou complexidade) do exemplo da figura 15.

As obras de referência publicadas são muito úteis para o indexador, principalmente na definição do significado de termos pouco comuns. Particularmente importantes são os dicionários e enciclopédias especializados e gerais, bem como os glossários de todos os tipos. Bakewell (1987) elaborou uma lista de obras de referência de interesse potencial para o indexador, porém hoje ela se apresenta muito desatualizada. Em algumas instituições o trabalho do indexador conta com o auxílio do acesso em linha a bancos de dados terminológicos.

CAPÍTULO 4

Índices pré-coordenados

A flexibilidade inerente aos sistemas pós-coordenados deixa de existir quando os termos de indexação são impressos em papel ou fichas catalográficas convencionais. Os índices impressos e os catálogos em fichas são *pré-coordenados*; suas características são as seguintes:

1. É difícil representar a multidimensionalidade das relações entre os termos.
2. Os termos somente podem ser listados numa determinada seqüência (A, B, C, D, E), o que implica que o primeiro termo é mais importante do que os outros.
3. Não é fácil (senão completamente impossível) combinar termos no momento em que se faz uma busca.

A forma mais rudimentar de um sistema de recuperação da informação talvez seja o tradicional catálogo em fichas utilizado há séculos nas bibliotecas. Vejamos o item mencionado anteriormente: um livro sobre migração de mão-de-obra de Moçambique para as minas da África do Sul. Suponhamos que lhe tenham sido atribuídos três cabeçalhos de assuntos: MOÇAMBIQUE, ÁFRICA DO SUL e TRABALHADORES MIGRANTES. A descrição bibliográfica do livro apareceria sob todos os três cabeçalhos num catálogo alfabético de assuntos em formato de fichas. Isso faz com que se tenha acesso ao livro sob qualquer um desses cabeçalhos. Será, entretanto, extremamente difícil realizar uma busca a respeito de qualquer *combinação* desses termos. Por exemplo, um usuário que esteja procurando livros sobre as relações políticas ou econômicas entre Moçambique e África do Sul precisaria examinar todas as entradas sob o cabeçalho MOÇAMBIQUE ou sob o cabeçalho ÁFRICA DO SUL. Mesmo que o fizesse, não reconheceria necessariamente os itens pertinentes. Se procurasse sob MOÇAMBIQUE, provavelmente só reconheceria que um livro era pertinente se o mesmo contivesse em seu título o termo 'África do Sul' (e vice-versa, se procurasse sob ÁFRICA DO SUL), ou se no pé da ficha catalográfica aparecessem os outros cabeçalhos atribuídos ao livro (seria improvável que os consultasse, a menos que fosse um usuário de catálogos muito experiente). Outra possibilidade seria procurar sob todas as entradas com MOÇAMBIQUE e todas as entradas com ÁFRICA DO SUL para tentar encontrar títulos que ocorressem sob ambas — um processo muito enfadonho, se houver muitas entradas para consultar.

É possível melhorar essa situação nos catálogos em fichas mediante o emprego de um cabeçalho como subcabeçalho (isto é, os termos são *pré-coordenados* numa entrada). Assim, ter-se-ia uma entrada como a seguinte:

Moçambique – Relações Econômicas

ou mesmo

Moçambique – Relações Econômicas – África do Sul

Os subcabeçalhos, no entanto, costumam ser adotados de maneira relativamente parcimoniosa nos catálogos de bibliotecas, e seria raro o catálogo que reunisse toda uma seqüência* de termos como na seguinte entrada pré-coordenada:

Moçambique, Relações Econômicas, África do Sul, Trabalhadores Migrantes, Mineiros

É mais provável que entradas detalhadas como essa apareçam em índices impressos do que em catálogos em fichas. A este respeito, os índices impressos são considerados ferramentas de recuperação mais eficientes do que os catálogos convencionais de bibliotecas. Em certos índices impressos, o usuário percorreria as entradas sob Moçambique para verificar se alguma delas também menciona a África do Sul. Exemplos de várias formas de índices impressos encontram-se no capítulo 10.

Mas uma entrada como essa do exemplo apresenta um problema óbvio: ela proporciona acesso ao documento somente para quem estiver procurando sob o termo MOÇAMBIQUE, sem dar acesso numa busca relativa à África do Sul, mineiros ou trabalhadores migrantes. Para que sejam oferecidos pontos de acesso adicionais é preciso criar mais entradas no índice.

Não existe maneira alguma pela qual um índice impresso possa proporcionar, de forma econômica, o mesmo nível de acesso ao documento que é proporcionado por um sistema de recuperação pós-coordenado. Conforme mostramos anteriormente, um sistema pós-coordenado permite o acesso por meio de qualquer *combinação* de termos atribuídos ao documento. O número de combinações é 2^n-1, onde n representa o número de termos. Assim, para um item indexado sob cinco termos, haverá 2^5-1 combinações, ou seja, um total de 31. Teoricamente, então, um índice impresso proporcionaria todas as combinações de cinco termos, se imprimisse 31 entradas. Seria economicamente inviável criar um índice impresso que contivesse tantas entradas para cada item, e a quantidade de entradas aumentaria dramaticamente à medida que aumentasse o número de termos — existem 255 combinações de oito termos!

Além do mais, como os termos devem ser impressos um em seguida ao outro numa entrada (isto é, numa seqüência linear), aos índices impressos preside a *permutação* e não a *combinação.* Por exemplo, a seqüência MOÇAMBIQUE, ÁFRICA DO SUL não é a mesma de ÁFRICA DO SUL, MOÇAMBIQUE. O número de permutações é n fatorial, sendo n o número de termos. Por exemplo, o número de permutações de oito termos é 40 320 ($8\times7\times6\times5\times4\times3\times2\times1$).

A situação dos índices impressos não é tão desoladora quanto essas conside-

---

* Por isso, esse tipo de indexação é às vezes chamado *indexação em seqüência* [*string indexing*] (Craven, 1986).

rações dão a entender. Vários programas de computador foram desenvolvidos para gerar automaticamente um conjunto de entradas de índice a partir de uma seqüência de termos. Um desses processos é conhecido como SLIC (Selective Listing in Combination [Listagem Seletiva em Combinação]. O programa, criado por Sharp (1966), primeiro organiza a seqüência de termos em ordem alfabética. Esta seqüência (ver figura 17) torna-se a primeira entrada do índice. O programa gera, então, todas as demais entradas julgadas necessárias, obedecendo a duas regras simples:

1. Os termos são sempre listados em ordem alfabética.
2. As seqüências redundantes são eliminadas (por exemplo, a entrada Trabalhadores Migrantes, Mineiros não será necessária se já houver Trabalhadores Migrantes, Mineiros, África do Sul).

Quando esta regra é obedecida, a quantidade de entradas cai de 2^n-1 para 2^{n-1}.

África do Sul
Mineiros, África do Sul
Mineiros, África do Sul, Moçambique
Relações Econômicas, África do Sul
Relações Econômicas, África do Sul, Mineiros
Relações Econômicas, África do Sul, Mineiros, Moçambique
Relações Econômicas, África do Sul, Mineiros, Moçambique, Trabalhadores Migrantes
Relações Econômicas, África do Sul, Mineiros, Trabalhadores Migrantes
Relações Econômicas, África do Sul, Moçambique
Relações Econômicas, África do Sul, Trabalhadores Migrantes
Relações Econômicas, Trabalhadores Migrantes, Moçambique, África do Sul,
Trabalhadores Migrantes, África do Sul
Trabalhadores Migrantes, África do Sul, Mineiros
Trabalhadores Migrantes, África do Sul, Mineiros, Moçambique
Trabalhadores Migrantes, África do Sul, Moçambique

FIGURA 17
Entradas de um índice SLIC

O método SLIC é engenhoso, pois permite todas as justaposições úteis de termos, pelo menos enquanto estes forem mantidos em ordem alfabética. Mas também tem suas desvantagens: ainda gera um número bastante grande de entradas; o consulente, para usar o índice com eficiência, deve reorganizar mentalmente os termos de busca em ordem alfabética (por exemplo, encontrará Trabalhadores Migrantes, Moçambique, mas não Moçambique, Trabalhadores Migrantes); perde o contexto para os termos situados perto do início da ordem alfabética (por exemplo, quem procurasse todas as entradas sob África do Sul não teria idéia alguma sobre o assunto deste item).

Outros índices baseiam-se num conjunto de entradas que se obtêm sistematicamente mediante alternação [*cycling*], rotação ou deslocamento [*shunting*]. Na *alternação,* cada termo numa seqüência é movido para a posição mais à

esquerda, a fim de se tornar um ponto de entrada, sendo os demais termos listados depois dele:

ABCDE
BCDEA
CDEAB
DEABC
EABCD

Note-se que, após o termo de entrada, vêm primeiro os termos que o seguiam na seqüência original e, depois, os que originalmente o precediam. No índice alternado, a sucessão de termos numa seqüência não precisa dispor-se segundo uma ordem evidente, embora estejam freqüentemente ordenados alfabeticamente e possam ser ordenados 'sistematicamente' (como se verá adiante).

A *rotação* é essencialmente o mesmo que a alternação, exceto que o termo de entrada é ressaltado de alguma forma (por exemplo, grifado ou sublinhado), em vez de ser deslocado para a posição mais à esquerda:

*A*BCDE
A*B*CDE
AB*C*DE
ABC*D*E
ABCD*E*

Tanto a alternação quanto a rotação proporcionam um certo 'contexto' para um termo, mas as relações entre alguns dos termos ainda permanecem obscuras ou ambíguas. Um índice baseado no *deslocamento* emprega uma apresentação em duas linhas na tentativa de reduzir a ambigüidade (isto é, ser mais preciso ao mostrar como um termo se relaciona com outro), como nos exemplos:

A	B.A
B.C.D	C.D

O principal exemplo disso, que é o PRECIS, será examinado mais adiante.*

Um método simples para produzir um índice impresso, baseado na ordem alfabética e na 'alternação' sistemática de termos para que ocupem a posição de entrada, conforme utilizado nas séries da *Excerpta Medica,* encontra-se exemplificado na figura 18. Mais uma vez, a primeira entrada resulta da colocação de todos os termos em ordem alfabética. As entradas adicionais derivam da movimentação de cada termo, sucessivamente, para a posição de entrada, sendo os demais termos listados depois dele (sempre em ordem alfabética) como uma seqüência de modificadores. Ainda que isso não enseje todas as justaposições possíveis de termos, na realidade oferece algumas vantagens evidentes em comparação com o SLIC: é mais econômico (não há mais entradas do que a quan-

*A terminologia relativa a índices pré-coordenados não se acha realmente padronizada. Por exemplo, Craven (1986) parece que não faz distinção entre alternação e rotação.

tidade de termos atribuídos) e cada entrada conta com seu 'contexto' completo. Nesse tipo de índice impresso é possível reconhecer dois tipos de termos: os que geram entradas de índice e os que não as geram. Os termos que não irão gerar entradas são marcados de alguma forma pelo indexador (ou são reconhecidos automaticamente). Tais termos serão utilizados apenas como modificadores. Aparecem no final da seqüência de termos e são reconhecidos por estarem fora da ordem alfabética e talvez impressos com diferente estilo de letra (ver o exemplo 'bibliografia' na figura 18).

África do Sul, Mineiros, Moçambique, Relações Econômicas, Trabalhadores Migrantes, *Bibliografia*
Mineiros, África do Sul, Moçambique, Relações Econômicas, Trabalhadores Migrantes, *Bibliografia*
Moçambique, África do Sul, Mineiros, Relações Econômicas, Trabalhadores Migrantes, *Bibliografia*
Relações Econômicas, África do Sul, Mineiros, Moçambique, Trabalhadores Migrantes, *Bibliografia*
Trabalhadores Migrantes, África do Sul, Mineiros, Moçambique, Relações Econômicas, *Bibliografia*

FIGURA 18
Entradas de índice baseado na alternação sistemática (modelo da *Excerpta Medica*)

Os índices exemplificados nas figuras 17 e 18 pressupõem o emprego de termos de indexação e não de texto livre, embora, em princípio, possam ser produzidos por computador depois que, mediante programas, tenham sido extraídas do texto narrativo frases 'significativas'. Alguns métodos ainda mais simples de produção de índices impressos foram criados para trabalhar com textos e, especialmente, palavras que ocorrem nos títulos dos documentos. Os métodos mais adotados são o KWIC *(keyword in context)* [palavra-chave no contexto], KWOC *(keyword out of context)* [palavra-chave fora do contexto] e suas variantes.

O índice KWIC (Luhn, 1959) é um índice rotado, derivado, em sua forma mais comum, dos títulos de publicações. Cada *palavra-chave* que aparece num título torna-se ponto de entrada, destacada de alguma forma, aparecendo, normalmente, realçada no centro da página como no exemplo da figura 19. As palavras restantes do título aparecem 'envolvendo' a palavra-chave. O índice KWIC constitui o método mais simples de produção de índices impressos por computador, no entanto, tem alguma eficiência, pois cada palavra-chave é vista em seu 'contexto'. Por exemplo (figura 19); é possível percorrer as entradas para '*crystals*' [cristais] em busca das que pareçam tratar das propriedades elásticas ou plásticas dos cristais. Os índices KWIC normalmente remetem apenas para alguma forma de número de documento, sendo preciso reportar-se a esse número a fim de obter informações bibliográficas completas sobre o item representado.

Note-se que o programa de computador que gera o índice identifica as palavras-chave mediante um processo 'reverso': reconhece as que não são palavras-chave (constantes de uma lista de palavras proibidas) e impede que sejam adotadas como pontos de entrada. Os vocábulos dessa lista de palavras proibidas têm função sintática (artigos, preposições, conjunções, etc.), mas, em si mesmos,

não possuem conteúdo temático. O índice KWIC é um método barato de obter certo nível de acesso temático ao conteúdo de uma coleção. É útil na medida em que os títulos sejam bons indicadores de conteúdo (por isso, é provável que funcione melhor com certos assuntos ou tipos de materiais do que com outros), embora, em princípio, não haja motivo para que os índices KWIC não derivem de outro texto, como, por exemplo, frases de resumos ou até seqüências de cabeçalhos de assuntos. Muitos estudos foram feitos sobre a utilidade dos títulos na recuperação (ver Hodges, 1983, e Hjørland e Nielsen, 2001). Os títulos podem também ficar mais informativos com o *acréscimo* ou *enriquecimento*. Isto é, outras palavras são acrescentadas ao título, normalmente entre parênteses, para explicá-lo ou torná-lo uma descrição mais completa do conteúdo do item.

```
LE TECHNIQUE FOR THE STUDY OF THE ELASTICITY OF CRYSTALS.                                  A SIMP
            STRUCTURAL IMPERFECTIONS IN QUARTZ CRYSTALS.
    LINEAR COMPRESSIBILITY OF FOURTEEN NATURAL CRYSTALS.
 THE LINEAR COMPRESSIBILITY OF THIRTEEN NATURAL CRYSTALS.
                TRANSLATION GLIDING IN CRYSTALS.
                          TWINNED CRYSTALS.
          BENDING CREEP OF ICE SINGLE CRYSTALS.
  DIRECT MEASUREMENTS OF THE SURFACE ENERGY OF CRYSTALS.
      THE GROWTH AND DEFORMATION OF ICE CRYSTALS.
PRELIMINARY EXPERIMENTS ON THE PLASTICITY OF ICE CRYSTALS.                                 RESULTS OF P
          PROPAGATION OF CLEAVAGE CRACKS IN CRYSTALS.
S, IN DISLOCATIONS AND MECHANICAL PROPERTIES OF CRYSTALS.    THE DIRECT OBSERVATION OF DISLOCATION PAT
AL GRAINS, PETROFABRIC AND INTERFACE STRUCTURE, CRYSTALS.    THE ELASTIC CONSTANTS OF ROCKS IN TERMS O
    DISLOCATIONS AND MECHANICAL PROPERTIES OF CRYSTALS.  TEXTBOOK.
                DISLOCATIONS IN CRYSTALS. TEXTBOOK.
            PHYSICAL PROPERTIES OF CRYSTALS.TEXTBOOK.
                STRENGTH OF CRYSTALS.TEXTBOOK.IN GERMAN.
              PLASTICITY OF CRYSTALS.TEXTBOOK.
      IMPERFECTIONS IN NEARLY PERFECT CRYSTALS.TEXTBOOK.
    DISLOCATION AND PLASTIC FLOW IN CRYSTALS.TEXTBOOK.


      ANNEALING RECRYSTALLIZATION IN CALCITE CRYSTALS AND AGGREGATES.
BRIES.SEMINAR.      FLOW OF ROCK FORMING CRYSTALS AND AGGREGATES. KINK BANDS.  MINERALS.PETROFA
THE EFFECT OF ORIENTATION ON STRESSES IN SINGLE CRYSTALS AND OF RANDOM ORIENTATION ON STRENGTH OF POLY
            THE FAILURE OF CAVITIES IN CRYSTALS AND ROCKS UNDER PRESSURE.
   EXPERIMENTAL DEFORMATION OF QUARTZ SINGLE CRYSTALS AT 27-30 KB, CONFINING PRESSURE AT 24 DEG.C.
EN.          DEFORMATION OF CRYSTALS AT HIGH PRESSURE AND HIGH TEMPERATURE.IN GERM
.IN RUSSIAN.    OBSERVATION OF DISLOCATIONS IN CRYSTALS BY THE METHOD OF SELECTIVE ETCHING.CONFERENCE
RY ANALYSIS OF PREFERRED ORIENTATION OF QUARTZ CRYSTALS IN 3 LINEATED QUARTZITE.                 T-
 ICE CRYSTALS IN GLACIERS COMPARED WITH QUARTZ CRYSTALS IN DYNAMICALLY METAMORPHOSED SANDSTONES.
DYNAMICALLY METAMORPHOSED SANDSTONES.    ICE CRYSTALS IN GLACIERS COMPARED WITH QUARTZ CRYSTALS IN
SO GEOTHERMOMETRIC CONSIDERATIONS ON THE QUARTZ CRYSTALS IN ANYOLITE OF THE ROSITA MONTANA. IN RUMANIAN
METHOD IN STUDY OF STATISTICS IN ORIENTATION OF CRYSTALS IN ROCKS AND ORES.                DIFFRACTION
AL.      PREFERRED ORIENTATION OF OLIVINE CRYSTALS IN TROCTOLITE OF THE WICHITA MOUNTAINS,OKLAHO
S.    COMPRESSIONAL WAVE VELOCITIES IN SINGLE CRYSTALS OF ALKALI FELDSPAR AT PRESSURES TO 10 KILOBAR
            ELASTIC PROPERTIES OF SINGLE CRYSTALS OF ANHYDRITE.
ON THE INHOMOGENEITY OF PLASTIC DEFORMATION IN CRYSTALS OF AN AGGREGATE.
            THE DEFORMATION OF SINGLE CRYSTALS OF ICE.CONFERENCE.
        MECHANICAL PROPERTIES OF SINGLE CRYSTALS OF ICE
                CREEP OF SINGLE CRYSTALS OF ICE.
SURE ON VELOCITY OF SHEAR DEFORMATION OF SINGLE CRYSTALS OF ICE.          EFFECT OF HYDROSTATIC PRES
            PLASTIC DEFORMATION OF SINGLE CRYSTALS OF QUARTZ.
```

FIGURA 19
Exemplo de entradas de um índice KWIC
Reproduzido de *KWIC' Index of Rock Mechanics Literature*, com permissão do American Institute of Mining, Metallurgical and Petroleum Engineers, Inc.

O índice KWOC é similar ao KWIC, exceto que as palavras-chave que se tornam pontos de acesso são repetidas fora do contexto, comumente destacadas na margem esquerda da página (figura 20) ou usadas como se fossem cabeçalhos de assuntos (figura 21). Faz-se às vezes uma diferença entre índices KWOC e índices KWAC (*keyword and context* [palavra-chave e contexto]). Quem adota essa distinção chama de índices KWAC os índices mostrados nas figuras 20 e 21. Um índice KWOC seria então aquele em que a palavra-chave usada como ponto de entrada não se repete no título mas é substituída por um asterisco (*) ou outro

símbolo. É difícil justificar essa prática insólita (empregar um símbolo para substituir a palavra-chave), de modo que a distinção entre KWOC e KWAC não é muito útil. Há diversas variantes de KWIC/KWOC, inclusive o KWIC duplo (Petrarca & Lay, 1969). Afins à família KWIC/KWOC são os índices de 'termo permutado', mais bem exemplificados pelo índice Permuterm, relacionado aos índices de citações produzidos pelo Institute for Scientific Information. No índice Permuterm cada palavra-chave do título é ligada, uma por vez, com outra palavra-chave nesse título, por exemplo:

CRISTAIS

ALUMÍNIO	20071
ANÁLISE	18024
COBALTO	00409
CRESCIMENTO	20071
DESLOCAÇÕES	04778
EQUILÍBRIO	17853
FERRITE	04778
HEXAGONAIS	30714

Com esse tipo de índice é fácil associar palavras-chave durante a busca, ao percorrer, por exemplo, a coluna de 'cristais' para verificar se algum dos títulos pode tratar de cristais de cobalto. Note-se que todas as palavras-chave do título aparecem reunidas em pares (por exemplo, o documento que tem em comum o número 04778 indica que os termos 'cristais', 'deslocações' e 'ferrite' ocorrem no mesmo título) e cada palavra-chave torna-se ponto de entrada no índice: 'alumínio' será ponto de entrada, e também 'análise', 'equilíbrio' e assim por diante.

De certo modo afim ao grupo de índices KWIC/KWOC/permutado tem-se o 'índice articulado de assuntos', exemplificado pelo índice de assuntos do *Chemical Abstracts*. Este tipo de índice usa uma breve descrição narrativa do documento para gerar as entradas. Esta descrição pode ser um enunciado redigido pelo indexador ou, em seu lugar, um título ou frase extraída do texto. Certas palavras ou frases que aparecem nesse enunciado são selecionadas como pontos de entrada no índice, mantendo-se o restante do enunciado como um modificador que proporciona o contexto necessário.

Armstrong e Keen (1982) descrevem o processo de elaboração de entradas para um índice articulado da seguinte forma:

> Os termos de entrada são reordenados de tal modo que cada um deles se liga a seu vizinho original por meio de uma palavra funcional ou pontuação especial, conservando-se assim a estrutura similar à de uma frase, ainda que muitas vezes disposta em ordem diferente (p.6).

Os seguintes exemplos, extraídos de Armstrong e Keen, demonstram o princípio:

> Indexação de Periódicos de Química por Pesquisadores
> Periódicos de Química, Indexação de, por pesquisadores
> Química, Periódicos de, Indexação de, por pesquisadores


```
NONEQUILIBRIUM  SCALE EFFECTS FOR NONEQUILIBRIUM CONVECTIVE HE
                AT TRANSFER WITH SIMULTANEOUS GAS PHASE AND SU
                RFACE CHEMICAL REACTIONS. APPLICATION TO HYPER
                SONIC FLIGHT AT HIGH ALTITUDES
                                AD-291 032(K)  $1.60 0025
NONLINEAR       APPLICATION OF VARIATIONAL EQUATION OF MOTION
                TO THE NONLINEAR VIBRATION ANALYSIS OF HOMOGEN
                EOUS AND LAYERED PLATES AND SHELLS
                                AD-289 868(K)  $2.60 0667
NONLINEAR       EXTENSIONS IN THE SYNTHESIS OF TIME OPTIMAL OR
                 BANG-BANG NONLINEAR CONTROL SYSTEMS. PART I.
                THE SYNTHESIS OF QUASI-STATIONARY OPTIMUM NONL
                INEAR CONTROL SYSTEMS
                                PB 162 547(K)  $4.60 0235
NONLINEAR       EXTENSIONS IN THE SYNTHESIS OF TIME OPTIMAL OR
                 BANG-BANG NONLINEAR CONTROL SYSTEMS. PART I.
                THE SYNTHESIS OF QUASI-STATIONARY OPTIMUM NONL
                INEAR CONTROL SYSTEMS
                                PB 162 547(K)  $4.60 0235
NONLINEAR       NONLINEAR FLEXURAL VIBRATIONS OF SANDWICH PLAT
                ES             AD-289 871(K)  $2.60 0669
NONLINEAR       OPTIMUM NONLINEAR CONTROL FOR ARBITRARY DISTUR
                BANCES         NASA N62-15890(K)  $2.60 0682
NONRECURRENT    A TECHNIQUE FOR NARROW-BAND TELEMETRY OF NONRE
                CURRENT PULSES AD-290 697(K)  $2.60 0577
NONUNIFORM      ELECTROMAGNETIC SCATTERING FROM A SPHERICAL NO
                NUNIFORM MEDIUM. PART II. THE RADAR CROSS SECT
                ION OF A FLARE AD-289 615(K)  $2.60 0747
NONUNIFORM      ELECTROMAGNETIC SCATTERING FROM ASPHERICAL NON
                UNIFORM MEDIUM. PART I. GENERAL THEORY
                                AD-289 614(K)  $2.60 0748
NORMAL          PROBABILITY INTEGRALS OF MULTIVARIATE NORMAL A
                ND MULTIVARIATE-T AD-290 746(K)  $8.60 0760
NORMAL          RESONANCE ABSORPTION OF GAMMA-RAYS IN NORMAL A
                ND SUPERCONDUCTING TIN
                                AD-289 844(K)  $3.60 0826
NORMS           NORMS FOR ARTIFICIAL LIGHTING
                                AD-290 555(K)  $1.10 0734
NORTH           FACTORS INFLUENCING VASCULAR PLANT ZONATION IN
                 NORTH CAROLINA SALTMARSHES
                                AD-290 938(K)  $7.60 0603
NORTH           SONAR STUDIES OF THE DEEP SCATTERING LAYER IN
                THE NORTH PACIFIC PB 162 427(K)  $2.60 0587
NORTH           THE DEVELOPMENT OF RESCUE AND SURVIVAL TECHNIQ
                UES IN THE NORTH AMERICAN ARCTIC
                                PB 162 410(K) $12.00 0085
NOSE            THE FLORA OF HEALTHY DOGS. I. BACTERIA AND FUN
                GI OF THE NOSE, THROAT, AND LOWER INTESTINE
                                LF-2(K)  $2.60 0458
NOZZLE          FABRICATION OF PYROLYTIC GRAPHITE ROCKET NOZZL
                E COMPONENTS   PB 162 371(K)  $1.10 0351
NOZZLE          FABRICATION OF PYROLYTIC GRAPHITE ROCKET NOZZL
                E COMPONENTS   PB 162 370(K)  $1.10 0353
NOZZLE          FABRICATION OF PYROLYTIC GRAPHITE ROCKET NOZZL
                E COMPONENTS   PB 162 372(K)  $2.60 0352
NOZZLE          THIRD SYMPOSIUM ON ADVANCED PROPULSION CONCEPT
                S SPONSORED BY UNITED STATES AIR FORCE OFFICE
                OF SCIENTIFIC RESEARCH AND THE GENERAL ELECTRI
                C COMPANY FLIGHT PROPULSION DIVISION CINCINNAT
                I, OHIO OCTOBER 2-4, 1962. PLASMA FLOW IN A MA
                GNETIC ARC NOZZLE AD-290 082(K)  $2.60 0147
NOZZLES         HEAT TRANSFER AND PARTICLE TRAJECTORIES IN SOL
                ID-ROCKET NOZZLES AD-289 681(K)  $5.60 0030
```

FIGURA 20

Amostra das entradas de um índice KWOC

Reproduzida de *U.S. Government Technical Reports*, Volume 1, 1963, com permissão do National Technical Information Service

Note-se que é mantida a sintaxe do texto original de modo que o significado do enunciado original não fica obscuro. Esses enunciados de indexação podem ser preparados pelo indexador, obedecendo a um conjunto prescrito de regras, ou podem ser desenvolvidos programas de computador que geram essas entradas (Armitage & Lynch, 1968; Lynch & Petrie, 1973).

GLYCIDE

MATERNAL GLYCIDE NORMAL ASSIMILATION. TOMATO BABY, PRECEDENTS OF MACROSOMIA AND FETAL MORTALITY. • B SALVADORI, G CAGNAZZO, A DELEONARDIS • MINERVA PEDIAT V12 P117, 11 FEB 60 IT

GLYCINE

AN INSULIN ASSAY BASED ON THE INCORPORATION OF LABELLED GLYCINE INTO PROTEIN OF ISOLATED RAT DIAPHRAGM. • K L MANCHESTER, P J RANDLE, F G YOUNG • J ENDOCR V19 P259-62, DEC 59

MAINTENANCE OF CARBOHYDRATE STORES DURING STRESS OF COLD AND FATIGUE IN RATS PREFED DIETS CONTAINING ADDED GLYCINE. • W R TODD, M ALLEN • USAF ARCTIC AEROMED LAB TECHN REP V57-34 P1-16, JUNE 60

GLYCINE C14

RATE OF ASSOCIATION OF S35 AND C14 IN PLASMA PROTEIN FRACTIONS AFTER ADMINISTRATION OF NA2S3504, GLYCINE-C14, OR GLUCOSE C14. • J E RICHMOND • J BIOL CHEM V234 P2713-6, OCT 59

GLYCOGEN

GLYCOGEN OF THE ADRENAL CORTEX AND MEDULLA. INFLUENCE OF AGE AND SEX. • M PLANEL, A GUILHEM • C R SOC BIOL PAR V153 P844-8, 1959 FR

EFFECT OF DIET ON THE BLOOD SUGAR AND LIVER GLYCOGEN LEVEL OF NORMAL AND ADRENALECTOMIZED MICE. • B P BLOCK, G S COX • NATURE LOND V184 SUPPL 10 P721-2, 29 AUG 59

LIVER GLYCOGEN AND BLOOD SUGAR LEVELS IN ADRENAL-DEMEDULLATED AND ADRENALECTOMIZED RATS AFTER A SINGLE DOSE OF GROWTH HORMONE. • C A DE GROOT • ACTA PHYSIOL PHARMACOL NEERL V9 P107-20, MAY 60

A MICROMETHOD FOR SIMULTANEOUS DETERMINATION OF GLUCOSE AND KETONE BODIES IN BLOOD AND GLYCOGEN AND KETONE BODIES IN LIVER. • O HANSEN • SCAND J CLIN LAB INVEST V12 P18-24, 1960

AN INVERSE RELATION BETWEEN THE LIVER GLYCOGEN AND THE BLOOD GLUCOSE IN THE RAT ADAPTED TO A FAT DIET. • P A MAYES • NATURE LOND V187 P325-6, 23 JULY 60

LIVER GLUCOSYL OLIGOSACCHARIDES AND GLYCOGEN CARBON-14 DIOXIDE EXPERIMENTS WITH HYDROCORTISONE. • H G SIE, J ASHMORE, R MAHLER, W H FISHMAN • NATURE LOND V184 P1380-1, 31 OCT 59

STUDIES ON GLYCOGEN BIOSYNTHESIS IN GUINEA PIG CORNEA BY MEANS OF GLUCOSE LABELED WITH C14. • R PHAUS, J OBENBERGER, J VOTOCKOVA • CESK FYSIOL V9 P45-6, JAN 60 CZ

GLYCOGEN CONTENT AND CARBOHYDRATE METABOLISM OF THE LEUKOCYTES IN DIABETES MELLITUS. • G MAEHR • WIEN Z INN MED V40 P330-4, SEPT 59 GER

GLYCOGEN LIVER. AN IATROGENIC ACUTE ABDOMINAL DISORDER IN DIABETES MELLITUS. • A SCHOTTE, H K LANKAMP, M FRENKEL • NED T GENEESK V103 P2258-62, 7 NOV 59 DU!

ACUTE GLYCOGEN INFILTRATION OF THE LIVER IN DIABETES MELLITUS. 2. THE EFFECTS OF GLUCAGON THERAPY. • A SCHOTTE, H K LANKAMP, M FRENKEL • NED T GENEESK V104 P1286-91, 2 JULY 60 DU!

FIGURA 21

Formato alternativo de um índice KWOC usado no *Diabetes-Related Literature Index*, suplemento de *Diabetes*, volume 12, 1960

Um exemplo de índice articulado de assuntos, que é, de fato, o mesmo descrito minuciosamente por Armstrong e Keen (1982), é o NEPHIS (Nested Phrase Indexing System [Sistema de Indexação de Frase Encaixada]), criado por

Craven (1977). Em sua forma mais simples, o indexador emprega colchetes angulares para indicar uma frase 'encaixada' numa frase maior e que será usada para gerar entradas de índice. Por exemplo, a frase

Produtividade das Pesquisas de <Especialistas do Sono>

gerará as duas entradas seguintes:

Produtividade das Pesquisas de Especialistas do Sono
Especialistas do Sono, Produtividade das Pesquisas de

Craven elabora este princípio simples com o acréscimo de outros símbolos e convenções a serem utilizados pelo indexador para criar entradas de índice que sejam coerentes e inequívocas, além de úteis. O trabalho de Armstrong e Keen (1982) nos dá uma idéia das possibilidades deste método de indexação relativamente simples. Bastante semelhante ao NEPHIS é o sistema PASI (Pragmatic Approach to Subject Indexing [Método Pragmático de Indexação de Assuntos]) descrito por Dutta e Sinha (1984).

Vale a pena citar brevemente outro sistema de indexação. O SPINDEX (Selective Permutation Index [Índice de Permutação Seletiva]), criado para a indexação de fundos arquivísticos, originalmente não passava de um índice KWAC ou KWOC (Burke, 1967). Em versões posteriores, sofreu alterações para produzir entradas de índice de dois níveis, que consistiam em palavras-chave principais e qualificadoras, como nos exemplos ARIZONA, Questões indígenas, e QUESTÕES INDÍGENAS, Arizona (Cook, 1980). Lamentavelmente, a sigla SPINDEX, com o significado de Subject Profile Index [Índice de Perfil de Assuntos], foi também usada para um formato diferente por parte de produtores de vários índices impressos, inclusive o American Bibliographical Center (que edita *Historical Abstracts* e *America: History and Life*). Este método, depois denominado ABC-SPINDEX (American Bibliographical Center's Subject Profile Index) para diferenciá-lo do SPINDEX, com o qual não tem relação, parece ser praticamente idêntico aos índices alternados utilizados pela *Excerpta Medica* (Falk & Baser, 1980).

Classificação em índices de assuntos

Todos os índices até aqui examinados adotam métodos que são 'alfabéticos', mas não 'sistemáticos'. Outros tipos de índices exigem que as entradas sejam construídas segundo princípios 'lógicos'. Esses métodos remontam a Cutter (1876), que estabeleceu regras para questões como entrada direta *versus* entrada invertida (História da Antiguidade ou Antiguidade, História?). Kaiser (1911) introduziu um enfoque mais elaborado, que reconhecia três categorias de termos: concretos, processos e termos de localidade. 'Concretos' são termos relativos a 'coisas', reais ou imaginárias, e 'processos' abrangem atividades. Kaiser determinava que os 'enunciados' de indexação apresentassem os termos em seqüência sistemática e não em ordem alfabética. Só eram permitidas três seqüências:

1. Concreto–Processo (como em Tubos–Soldagem ou Tubos de Aço–Soldagem)
2. Localidade–Processo (como em Argentina–Comércio)
3. Concreto–Localidade–Processo (como em Café–Brasil–Exportação)

A fim de obedecer às regras de Kaiser, o indexador deveria evidenciar um termo concreto que se achasse implícito. Por exemplo, o termo *dessalinização* tornar-se-ia Água–Dessalinização.

Atribui-se a Ranganathan o mais importante desenvolvimento que teve lugar depois disso. Embora seu nome esteja fundamentalmente ligado às teorias da classificação e a seu próprio esquema de classificação bibliográfica, a *Colon Classification* [Classificação dos Dois Pontos], Ranganathan também prestou importante contribuição à prática moderna da indexação alfabética de assuntos. Sua *indexação em cadeia* constitui uma tentativa de obter um processo de desenvolvimento coerente do índice alfabético de assuntos do catálogo sistemático (em forma de fichas ou de livro). Os princípios de seu esquema de classificação, bem como suas teorias da classificação, fogem ao escopo deste livro. Bastaria dizer que uma das principais características dos esquemas de classificação elaborados de conformidade com os princípios de Ranganathan é a 'síntese' ou 'construção de números'. Quer dizer, o número de classificação que representa um assunto complexo é obtido pela reunião dos elementos notacionais que representam assuntos mais simples. Por exemplo, o tópico 'confecção de roupas de lã na Alemanha no século XIX' é representado pela notação *AbCfHYqZh*, na qual *Ab* representa 'roupas', *Cf* 'lã', *H* 'confecção', *Yq* 'Alemanha', e *Zh* 'século XIX', sendo todos estes elementos notacionais retirados de diferentes partes do esquema de classificação e combinados numa seqüência ('ordem preferida' ou 'ordem de citação') especificada pelo compilador do esquema.

É óbvio que o índice alfabético de um catálogo sistemático elaborado segundo esses princípios deve ser desenvolvido de forma coerente, senão resultará em algo caótico e impossível de usar. A solução dada por Ranganathan a este problema — a indexação em cadeia — implica que se indexe cada degrau da cadeia hierárquica, do mais específico até o mais genérico. Assim, um item representado pela classificação *AbCfHYqZh* geraria as seguintes entradas no índice:

Século XIX, Alemanha, Confecção, Artigos de Lã, Roupas *AbCfHYqZb*
Alemanha, Confecção, Artigos de Lã, Roupas *AbCfHYq*
Confecção, Artigos de Lã, Roupas *AbCfH*
Artigos de Lã, Roupas *AbCf*
Roupas *Ab*

Evidentemente, o usuário desse tipo de índice deve fazer a busca obedecendo também a uma seqüência predefinida de termos. Por exemplo, se estivesse procurando informações sobre roupas na Alemanha no século XIX, de pouca valia lhe seria esse índice ao consultar o termo *roupas*.

Ao determinar a seqüência em que os números de classificação são combinados num esquema de classificação 'analítico-sintético' (freqüentemente denominado, um tanto equivocadamente, 'facetado'), Ranganathan chegou a cinco 'categorias fundamentais' e a uma fórmula de reuni-las. As categorias, Personalidade, Matéria, Energia, Espaço e Tempo, são combinadas nesta seqüência e a fórmula é às vezes denominada simplesmente 'PMEST' [onde o S corresponde à letra inicial de 'space', espaço em inglês].

O modo mais fácil de descrever a Personalidade é como 'a coisa em si'. Matéria é o material de que a coisa é composta. Energia é a ação realizada na ou pela coisa. Espaço é onde a ação se verifica, e Tempo é quando ela ocorre. A seqüência *AbCfHYqZh* obedece à ordem PMEST. Por conseguinte, a entrada num índice em cadeia de um item categorizado dessa forma será o inverso dessa ordem.

A seqüência 'lógica' das facetas estabelecida por Ranganathan para a construção de números pode ser também adotada em catálogos e índices alfabéticos de assuntos. Poder-se-ia, assim, elaborar uma entrada de índice lógica, de acordo com a fórmula PMEST, da seguinte forma:

Roupas: Artigos de Lã: Confecção: Alemanha:
Século XIX

Infelizmente, a fórmula PMEST é um pouco simplista. Ao indexar assuntos altamente complexos, é possível que uma categoria ocorra mais de uma vez (por exemplo, a tensão exercida sobre uma estrutura pode levar ao rachamento dessa estrutura, o que implica duas ocorrências diferentes da categoria 'energia'); algumas das categorias precisam ser subdivididas mais ainda (por exemplo, para indicar diferentes tipos de atividades); ademais, a fórmula PMEST não trata claramente de certos atributos que são importantes na indexação, tais como as *propriedades* dos materiais.

As teorias de Ranganathan, no entanto, tiveram profundo efeito nas práticas modernas da indexação de assuntos. Isso se verifica, de modo patente, na obra de Coates (1960), que postula um catálogo ou índice despido da rigidez dos cabeçalhos de assuntos preestabelecidos. Uma entrada de assunto deveria ser totalmente coextensiva com o conteúdo temático estudado, como no exemplo

Linhas de Transmissão de Eletricidade, Cabos Aéreos, Condutores,
Rompimento, Prevenção, Manutenção

Coates utiliza uma 'fórmula de importância' para estabelecer a seqüência em que os termos componentes serão reunidos. A seqüência básica que adota é Coisa, Parte, Material, Ação, Propriedade, a qual, porém, pode ser modificada em determinadas circunstâncias. O cabeçalho utilizado acima, por exemplo, adota a seqüência Coisa, Espécie, Parte, Ação, Agente. Os processos desenvolvidos por Coates foram adotados pelo *British Technology Index* (posteriormente denominado *Current Technology Index),* do qual ele foi o primeiro editor. A figura 22 mostra exemplos de entradas desse índice. Observe-se que um item aparece uma única vez no índice. Proporcionam-se acessos adicionais por meio de remissivas.

Pode-se também ponderar que as teorias de Ranganathan exerceram influência sobre o PRECIS (Preserved Context Index System [Sistema de Indexação de Contexto Preservado]), desenvolvido por Austin (Austin, 1984). No PRECIS, programas de computador geram um conjunto completo de entradas de índice e remissivas a partir de uma seqüência de termos e códigos de instrução fornecidos pelo indexador para cada item. O conteúdo temático de um documento é descrito por meio de uma série de termos colocados numa seqüência 'dependente do contexto'. Austin e Digger (1977) utilizam o seguinte exemplo:

Índia, Indústria algodoeira, Pessoal, Treinamento

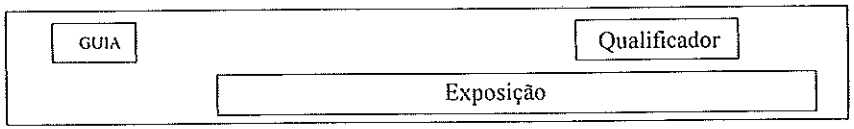
Alega-se que a lógica disso está em que cada termo é essencialmente dependente do termo que o antecede imediatamente. Assim, 'treinamento' aplica-se somente ao contexto de 'pessoal', 'pessoal' aplica-se somente ao contexto da 'indústria algodoeira', e esta se aplica somente ao contexto da Índia.

No PRECIS, as relações entre os termos componentes de uma entrada do índice são evidenciadas numa disposição em duas linhas:

Pessoal. Indústria algodoeira. Índia
Treinamento

Isso é justificado com o argumento de que proporciona uma forma prática de mostrar, simultaneamente, a relação entre o termo empregado como ponto de entrada no índice e os termos que são: a) de contexto mais amplo, e b) de contexto mais restrito. No exemplo acima, 'Pessoal' é modificado por 'Indústria algodoeira' e 'Índia' a fim de mostrar o contexto mais amplo, enquanto 'Treinamento' é apresentado como um dependente de 'Pessoal'.

Como se vê neste exemplo, uma entrada PRECIS contém três componentes:



O termo 'guia' é o ponto de entrada no índice, sendo impresso em negrito, o 'qualificador' apresenta o contexto mais amplo, e a 'exposição' mostra os termos de contexto mais restrito. Embora a posição de guia esteja evidentemente sempre ocupada, as outras posições nem sempre precisam estar ocupadas.

Entradas do tipo geral acima exemplificado podem ser geradas por computador a partir de uma série de termos apresentados numa seqüência dependente de contexto. Assim, a seqüência Índia, Indústria algodoeira, Pessoal, Treinamento geraria as seguintes entradas:

Índia
Indústria algodoeira. Pessoal. Treinamento
Indústria algodoeira. Índia
Pessoal. Treinamento

Infelizmente o processo não é tão simples quanto o que sugere este único exemplo. Há muitos casos em que a seqüência numa série de termos não revela, por si mesma e de modo inequívoco, as dependências. Na realidade, um indexador que utilize o PRECIS deverá empregar 'operadores' (códigos afixados aos termos componentes), a fim de representar sem ambigüidade as relações entre os termos. Para o exemplo mostrado antes a seqüência de entrada seria

(0) Índia
(1) indústria algodoeira
(P) pessoal
(2) treinamento

onde (2) representa 'ação de transição', (P) 'objeto de ação, parte do sistema-chave', (0) 'localização', e (1) 'sistema-chave' (objeto da ação transitiva). Estes operadores mostram a 'função' que um termo desempenha em relação a outros termos (proporcionando uma espécie de sintaxe) e assim podem ser considerados como 'indicadores de função' ou 'operadores de função'.

Austin e Digger apresentam uma lista de 26 operadores desse tipo. É claro que a utilização desse esquema torna bastante complicada a operação de indexação e eleva seu custo, além de exigir, para implantá-lo, um extenso manual de instruções para a indexação.

De algum modo relacionado com o PRECIS há o sistema POPSI (Postulate-based Permuted Subject Indexing [Indexação Permutada de Assunto com Base em Postulados]) (Bhattacharyya, 1981), inspirado nas teorias de Ranganathan sobre classificação.

O esquema de indexação de Farradane (1967, 1980), anterior ao PRECIS, guarda semelhança com este por também utilizar um esquema de indicadores de função. Enquanto o PRECIS emprega suas funções exclusivamente como meio de gerar por computador enunciados de indexação coerentes, as funções são reservadas no sistema de Farradane para indicar relações precisas entre os termos. Estas relações se baseiam nos trabalhos de psicologia experimental do raciocínio, de Piaget, Vinacke, Isaacs e outros autores, e corroboradas pelo trabalho de Guilford sobre a 'estrutura do intelecto'.

No esquema de Farradane há nove relações explícitas, cada uma representada por um 'operador'. O conjunto completo dos operadores encontra-se na figura 23. O esquema representa estádios de desenvolvimento do raciocínio extraídos da psicologia infantil, isto é, os estádios pelos quais a criança se desenvolve ao associar e diferençar objetos. Há dois conjuntos de gradação: em mecanismos associativos e em mecanismos discriminativos. O primeiro estádio associativo é a percepção simples sem referência ao tempo; o segundo é a associação temporária entre idéias; e o terceiro é a associação fixa (permanente) de idéias. Os estádios de discriminação são: coincidência simples (conceitos difíceis de discriminar), não-distinto (conceitos que têm muito em comum) e conceituação distinta (conceitos que podem ser completamente discriminados).

FABRICS
Related Headings:
WEAVING
FABRICS, Cellulosic,, Crease resistant : Cross linking ; Dimethylol-1,3-propylene urea
Deferred curing [BP 1,107,796: Sun Chemical Corp., USA] Dyer, Textile Printer, Bleacher & Finisher, 141 (2 May 69) p.614+
FABRICS, Cellulosic,, Crease resistant : Finishing
Crease-resist and wash-and-wear finishing. B.C.M. Dorset. Textile Manufacturer, 95 (Apr 69) p.156-63
FABRICS, Cellulosic,, Knitted,, Crease resistant : Finishing
Permanent press processes for knitted fabrics. D. Haigh. Hosiery Trade J., 76 (May 69) p.127+. il.
FABRICS; Cellulosic–Nylon : Dyeing, High temperature : Dyes, Reactive
"Hot-dyeing" reactive dyes on blends [Drimarene X and Drimafon X: Sandoz Products Ltd, Horsforth, Leeds] [summary] P.F. Bell. Dyer, Textile Printer, Bleacher & Finisher, 141 (2 May 69) p.622+
FABRICS; Cellulosic–Polyester fibres : Dyeing, High temperature : Dyes, Reactive
"Hot-dyeing" reactive dyes on blends [Drimarene X and Drimafon X: Sandoz Products Ltd, Horsforth, Leeds] [summary] P.F. Bell. Dyer, Textile Printer, Bleacher & Finisher, 141 (2 May 69) p.622+
FABRICS, Coated : Clothing. See CLOTHING : Fabrics, Coated
FABRICS : Finishing : Weft straighteners : Control system, Photoelectric
Fabric straightening [BP 1,107,822]H. Elcken. Dyer, Textile Printer, Bleacher & Finisher, 141 (2 May 69) p.612+
FABRICS, Foamback : Laminating
Versatility the key: different cloths call for different techniques. P. Lennox-Kerr. Hosiery Times, 42 (Apr 69) p.107-9. il.
FABRICS; Man made fibres,, Pile : Knitting
Manufacture and use-development of pile fabrics in Du Pont fibres. J. Rest & M.R.B. Addison. Hosiery Times, 42 (Apr 69) p.88+. il.
FABRICS; Mohair : Suitings. See SUITINGS : Fabrics; Mohair
FABRICS : Tape. See TAPE : Fabrics
FABRICS, Warp knit : Dyeing, High temperature : Heating : Heat transfer oil
HT process heating in the modern dyehouse [Kestner-Stone-Vapor at Nyla-Raywarp, Long Eaton] Dyer, Textile Printer, Bleacher & Finisher, 141 (18 Apr 69) p.542+
FABRICS, Warp knit : Knitwear. See KNITWEAR : Fabrics, Warp knit
FABRICS; Wool,, Knitted,, Shrink resistant : Finishing : Solvents : Perchloroethylene : Machines
'Bentley Rapide' solvent finishing machine for knitwear and piece-goods. A.G. Brooks. Hosiery Times, 42 (Apr 69) p.45+. il.
Milling machine for knitwear [Bentley Rapide] Hosiery Trade J., 76 (May 69) p.130+. il.

FIGURA 22
Exemplo de entradas do *British Technology Index*
Reproduzido com gentil permissão de CSA

Constróem-se os enunciados de indexação mediante a reunião de termos ('isolados'), usando esses operadores. Um enunciado de indexação, formado por termos relacionados entre si por meio de operadores, é denominado um 'analeto'. Eis alguns exemplos simples:

Aves /* Migração
Minério de ferro /–Fundição

e outro mais complexo:

Vidro/(Oxigênio/)Flúor/–Substituição

que representa a substituição do oxigênio pelo flúor no vidro. Utiliza-se uma apresentação bidimensional, quando necessário, como em:

Beterrabas/–Armazenamento /; Lavagem		O armazenamento de beterrabas lavadas
Rato /*	{Sucrose} {Óleo de coco}/–Alimentação	Ratos alimentados com sucrose com óleo de coco

Farradane (1977) comparou seu sistema de indexação relacional com o PRECIS, o NEPHIS e o POPSI, aos quais se refere de modo impreciso como capazes de produzir índices 'permutados'. Alega ser possível converter por computador seus diagramas bidimensionais em entradas de índices alfabéticos permutados.

		Mecanismos associativos		
		Percepção	Associação temporária	Associação fixa
Mecanismos discriminativos	Conceituação coincidente	1 /θ Coincidência	4 /* Auto-atividade	7 /; Associação
	Conceituação não-distinta	2 /= Equivalência	5 /+ Dimensional	8 /(Pertença
	Conceituação distinta	3 /) Distinção	6 /– Ação	9 /: Dependência funcional (causação)

FIGURA 23
Sistema de relações de Farradane
Reproduzido de Farradane (1980) com gentil permissão de CSA

Gardiner et al. (1995) reconhecem a influência de Farradane em sua abordagem das buscas em bases de dados de textos. Isto é, seus procedimentos de busca procuram textos em que os termos desejados parecem relacionar-se entre si na forma exata exigida pelo enunciado de busca.

O Symbolic Shorthand System [Sistema Taquigráfico Simbólico] (Selye, 1966; Selye e Ember, 1964) é outro sistema de indexação que expressa relações entre termos mediante indicadores de função. O indexador extrai os termos de um esquema de classificação, que compreende 20 classes principais, organizado predominantemente com base no sistema do corpo humano. Em todo o esquema são empregados símbolos mnemônicos [válidos para a língua inglesa, N.T.] para representar os assuntos. Por exemplo, *Adr* representa a glândula ad-renal, *Hypt* hipotálamo, *BMR* a taxa de metabolismo basal, e assim por diante. O principal indicador de função de Selye é uma flecha (←) que mostra a direção da ação, como em:

Cer ← ACTH	Efeito do hormônio adrenocorticotrófico sobre o cérebro

ou no exemplo mais complexo:

Adr ← Hyp← ACTH+TX	Efeito sobre a ad-renal da hipofisectomia em associação com o hormônio adrenocorticotrófico e a tiroxina

Outros indicadores de função mostram outras relações. Por exemplo, o símbolo < é usado para indicar conteúdo ou componente (Glu < B representa açúcar no sangue) e os dois pontos (:) para a função de comparação. Conteúdos temáticos bastante complexos podem ser representados de modo conciso e inequívoco neste sistema, conforme mostram os seguintes exemplos:

R ← ('B/Rb ← R/Duck')/Rat

(Injeção de substância renal do pato no sangue de coelho e injeção do soro assim obtido em ratos, produzindo alterações renais)

Glu < B (:Ur) ← CON

(Efeito da cortisona sobre o conteúdo de açúcar no sangue comparado com o conteúdo de açúcar na urina)

Nível de coordenação

Estabeleceu-se uma distinção entre sistemas pré-coordenados e pós-coordenados. Na realidade, porém, é provável que um sistema de recuperação da informação moderno incorpore características de pré-coordenação, bem como recursos de pós-coordenação. Possivelmente haverá certa pré-coordenação no vocabulário utilizado na indexação. Por exemplo, o descritor CRESCIMENTO POPULACIONAL, que se encontra em um tesauro, representa a pré-coordenação dos termos CRESCIMENTO e POPULAÇÃO. Em alguns sistemas, o indexador conta com a possibilidade de utilizar certos termos como subcabeçalhos de outros. Assim, ele pode criar:

CRESCIMENTO POPULACIONAL/ESTATÍSTICA

Finalmente, a pessoa que faz a busca pode combinar termos livremente em relações lógicas, como, por exemplo, 'recuperar itens indexados sob CRESCIMENTO POPULACIONAL/ESTATÍSTICA e também sob AMÉRICA DO SUL'.

Ocorre, então, uma certa coordenação (de conceitos ou termos que os representam) nas características do vocabulário, e mais alguma coordenação talvez ocorra no momento da indexação. Pode-se considerar isso como formas de *pré-coordenação,* uma vez que a coordenação está incorporada nos registros que dão entrada numa base de dados. O nível final de coordenação é aquele que se realiza por meio da manipulação de termos quando da realização de uma busca (isto é, *pós-coordenação).*

Embora este capítulo tenha apresentado exemplos de vários tipos de índices pré-coordenados, certamente não esgotou todas as possibilidades. Encontra-se

uma análise mais completa das características dos índices pré-coordenados em Keen (1977a) e Craven (1986). Keen (1977b) também examina o tema da estratégia de busca aplicada a esses índices.

Índices de final de livro

Ainda que muitos dos princípios examinados neste livro sejam válidos para índices de todos os tipos, sua atenção se concentra principalmente na indexação destinada a bases de dados de itens bibliográficos — indexação pós-coordenada para bases de dados em formato eletrônico, e indexação pré-coordenada para aquelas em forma impressa. Não se tentou apresentar instruções minuciosas sobre a indexação de livros como peças isoladas. Este tópico encontra-se bem estudado em outras publicações (por exemplo, Mulvany, 1994; *Guidelines for indexes*, 1997). Diodato (1994) apresenta resultados de estudo sobre preferências dos usuários em matéria de índices de livros; são comparadas as opiniões de bibliotecários e pessoal docente.

Os estudos mais completos sobre índices de livros parecem ser os relatados por Bishop et al. (1991) e Liddy et al. (1991). Nesse par de estudos afins, o primeiro analisa as características de uma amostra de índices (formato, arranjo e questões similares), enquanto o segundo examina as políticas das editoras (por exemplo, quem elabora o índice, exigências formais); este artigo também inclui algumas informações sobre características dos índices e conclusões relativas ao projeto como um todo. Liddy e Jörgensen (1993a) usaram estudantes como voluntários, a fim de verificar como realmente utilizavam o índice de um livro.

Índices pré-coordenados *versus* índices pós-coordenados

Os índices impressos do tipo que foi examinado neste capítulo podem ser muito eficazes na localização de um ou 'alguns' itens sobre um assunto de modo bem rápido. Alguns autores, porém, parecem exagerar ao louvar as virtudes dos índices pré-coordenados. Criticam a recuperação pós-coordenada com o argumento de que seus resultados são medíocres (ver Weinberg, 1995, por exemplo), como, por exemplo, excessiva irrelevância, embora isso possa ocorrer com todos os métodos, e que muitos usuários têm dificuldade para compreender a lógica das buscas. Essa última alegação é certamente verdadeira, mas também é verdade que muitas pessoas enfrentam enorme dificuldade para entender e usar o mais simples dos índices impressos (ver, por exemplo, Liddy e Jörgensen,1993a,b). Diante da opção, os usuários das bibliotecas parecem preferir, de modo esmagador, as buscas pós-coordenadas em bases de dados eletrônicas em comparação com o uso dos índices impressos (ver, por exemplo, Massey-Burzio, 1990), embora, de fato, possam obter resultados muito inferiores em suas buscas (ver p. 121-127 de Lancaster e Sandore, 1997).

CAPÍTULO 5

Coerência da indexação

É mais do que evidente que a indexação é um processo subjetivo e não objetivo. Duas (ou mais) pessoas possivelmente divergirão a respeito do que trata uma publicação, quais aspectos merecem ser indexados, ou quais os termos que melhor descrevem os temas selecionados. Ademais, uma mesma pessoa decidirá de modo diferente quanto à indexação em momentos diferentes. A *coerência* na indexação refere-se à extensão com que há concordância quanto aos termos a serem usados para indexar o documento. A *coerência interindexadores* refere-se à concordância entre indexadores, enquanto a *coerência intra-indexador* refere-se à extensão com que um indexador é coerente consigo mesmo.

Já foram adotadas ou propostas várias medidas diferentes para a coerência, e sobre as quais existe uma boa revisão bibliográfica feita por Leonard (1975). Talvez a medida mais comum seja a simples relação $AB/(A+B)$, onde A representa os termos atribuídos pelo indexador *a*, B representa os termos atribuídos pelo indexador *b*, e AB representa os termos com os quais *a* e *b* concordam. Vejamos a situação retratada na figura 24. Cinco pessoas indexaram o mesmo item, com o número de termos atribuídos variando de quatro (indexador *b*) a oito (indexador *e*). Podem-se comparar os termos atribuídos por qualquer par de indexadores. Hooper (1965) refere-se aos valores da coerência de pares como *pares de coerência* (PCs). Para os indexadores *a* e *b*, o PC é 3/6 ou 0,5 (existem seis termos exclusivos atribuídos e três deles foram atribuídos por ambos). Cada par do grupo é tratado da mesma forma. A partir dos dados apresentados são derivados os seguintes PCs: *ab*, (0,5); *ac*, 4/7 (0,57); *ad*, 4/6 (0,75); *ae*, 4/9 (0,44); *bc*, 3/7 (0,43); *bd*, 2/7 (0,29); *be*, 4/8 (0,5); *cd*, 3/8 (0,37); *ce*, 5/9 (0,56); *de*, 3/10 (0,30).

Obtém-se uma medida da coerência intergrupal por meio da determinação da média dos resultados para cada par de indexadores. Para o grupo *a–e* a coerência global é de aproximadamente 0,47.

Se a seqüência de termos na figura 24 reflete prioridade na atribuição de termos, verifica-se que existe concordância razoável quanto aos termos mais importantes. Todos os cinco indexadores atribuem o termo A, e quatro deles atribuem tanto A *quanto* B. Verifica-se muito menos concordância quanto aos aspectos secundários do documento ou quais os termos a serem atribuídos a esses aspectos. Observe-se também como a quantidade de termos atribuídos influi no escore da coerência: quanto mais termos atribuídos (pelo menos até certo ponto), menor tenderá a ser a coerência. Zunde e Dexter (1969b) e Rolling (1981) sugerem que as medidas de coerência deveriam levar em conta a importância de

diversos termos para o conteúdo temático do documento. A incoerência na atribuição de termos de menor importância será muito menos significativa do que a incoerência na atribuição de termos de maior importância, e isso se refletiria em qualquer método de pontuação.

a	*b*	*c*	*d*	*e*
A	*A*	*A*	*A*	*A*
B	*B*	*C*	*B*	*B*
C	*E*	*D*	*C*	*D*
D	*F*	*E*	*D*	*E*
E		*F*	*H*	*F*
		G		*G*
				I
				J

FIGURA 24

Termos (*A–J*) atribuídos ao mesmo documento por cinco indexadores diferentes (*a–e*)

Os dados da figura 24 poderiam também representar a coerência intra-indexador: a situação em que uma pessoa indexa o mesmo documento em cinco ocasiões diferentes.

Cooper (1969) considera a coerência interindexadores de modo diferente: no nível do termo. Quer dizer, ele mede o grau com que um grupo de indexadores concorda com a atribuição de determinado termo a um documento. Com relação a esse termo, a coerência interindexadores é definida como a proporção de indexadores que atribuem o termo menos a proporção daqueles que não o atribuem. No exemplo da figura 24 há 100% de concordância quanto ao termo A, enquanto a concordância quanto a B tem um valor de 60% (80%—20%), a concordância quanto a C tem um valor de 20 (60%—40%), e assim por diante.

Já foram realizados muitos estudos sobre coerência interindexadores, embora hoje não sejam tão comuns quanto no passado; eles costumam mostrar que é muito difícil alcançar alto nível de coerência. Hooper (1965) fez um resumo de 14 estudos diferentes e encontrou valores que variavam de 10% a 80%. Para os seis estudos em que pôde recalcular os valores a partir dos dados fornecidos (para ter certeza de que a coerência seria calculada da mesma forma para cada um), os resultados variaram de 24% a 80%.

Praticamente todos os estudos sobre coerência interindexadores até hoje realizados tratam cada termo como igual, embora, conforme sugerido antes, fosse mais sensato atribuir um 'peso' maior à coerência na atribuição dos termos mais importantes. Outra complicação está no fato de que, com certos tipos de vocabulários controlados e procedimentos de indexação, seria possível a ocorrência de uma coincidência parcial. Por exemplo, dois indexadores concordariam com o mesmo cabeçalho principal, mas não com o subcabeçalho. Vejamos o exemplo a seguir em que as letras maiúsculas representam cabeçalhos e o asterisco marca os cabeçalhos que o indexador considera mais importantes:

Indexador 1	Indexador 2	Indexador 3
*A/b	*A/c	*A/b
*B/b/c	*B/c	B/c
C/f	C/f	*D/f
D/f	D/r	F
E	F	*H/q
	G	I

Trata-se de uma situação realista. Por exemplo, ela se assemelha de perto à prática de indexação da National Library of Medicine onde mais de um subcabeçalho pode ser atribuído a um termo e os descritores principais são diferençados dos menos importantes.

É claro que esse tipo de indexação apresenta problemas importantes na realização de estudos de coerência. Aqui deixa de ter significado o método simples do par de coerência. Na indexação desse tipo, dever-se-ia dar mais 'crédito' a uma perfeita concordância entre dois indexadores. Por exemplo, os indexadores 1 e 3 mereceriam grande crédito pelo fato de ambos concordarem com a combinação A/b de cabeçalho principal/subcabeçalho e de que este seria um descritor principal. Embora seja possível desenvolver um método de pontuação numérica para expressar a coerência (5 pontos para uma perfeita concordância cabeçalho principal/subcabeçalho, 10 pontos para uma concordância de cabeçalho principal/subcabeçalho se ambos os indexadores o utilizarem como descritor mais importante, e assim por diante), é difícil chegar a um acordo sobre qual seria o escore, e mais difícil ainda interpretar o que o escore realmente significa. É mais provável que esse tipo de pontuação seja aplicável a estudos de *qualidade* de indexação, que é objeto do próximo capítulo.

Fatores que influem na coerência

Essa variabilidade nos escores da coerência leva a se indagar 'quais são os fatores que têm maior efeito na determinação da coerência na indexação?' Na figura 25 procuram-se identificar possíveis fatores.

Já se mencionou a quantidade de termos atribuídos. Se se pedisse aos indexadores que atribuíssem termos, em ordem de 'importância' percebida, ao conteúdo temático do documento, provavelmente obter-se-ia razoável grau de concordância no que concerne aos termos do alto da lista. Na medida em que se descer nessa lista, essa concordância certamente diminuirá. Em outras palavras, é certo que haverá mais concordância quanto aos tópicos do documento considerados principais do que quanto aos tópicos considerados de menor importância que mereçam ser incluídos.

Isso, porém, talvez seja um pouco simplista. A figura 26 sugere uma relação possível entre coerência e quantidade de termos atribuídos. Supondo que os termos sejam atribuídos em ordem de prioridade, levanta-se a hipótese de que a concordância atingirá o ponto mais alto no nível de dois termos e em seguida

começará a cair gradualmente até o ponto onde tenham sido atribuídos tantos termos que a concordância voltará a aumentar. Isto é exemplificado na figura 27.

1. Quantidade de termos atribuídos
2. Vocabulário controlado *versus* indexação com termos livres
3. Tamanho e especificidade do vocabulário
4. Características do conteúdo temático e sua terminologia
5. Fatores dependentes do indexador
6. Instrumentos de auxílio com que conta o indexador
7. Extensão do item a ser indexado

FIGURA 25
Possíveis fatores que influem na coerência da indexação

Essa figura apresenta listas ordenadas segundo a importância dos termos atribuídos pelos indexadores *a* e *b*. Isto é, *a* acha que *A* é o termo mais importante, *B* é o que se segue em ordem de importância, e assim por diante. Outra forma de examinar isso é dizer que, se o indexador *a* pudesse atribuir somente um termo ao documento, esse termo seria *A*. Cada indexador finalmente atribui 16 termos. Observe-se que, embora os indexadores concordem com os dois termos do alto da lista, eles não concordam com o primeiro desses termos. Isso não constitui surpresa. Muitos documentos envolvem uma relação entre dois conceitos principais. Talvez seja possível estar de acordo sobre quais são esses conceitos, mas não concordar com qual deles assumirá precedência. Por exemplo, num artigo sobre soldagem de titânio, é o metal ou o processo que deve assumir precedência? (É claro que decisões como essa têm muito a ver com as características da base de dados. Numa que seja dedicada exclusivamente ao titânio, o termo *titânio* tem pouco ou nenhum valor.) Isso se parece um pouco com apostar em cães (ou cavalos) de corrida: amiúde é mais fácil adivinhar quais os dois animais que terminarão nas primeiras duas posições do que adivinhar qual será o primeiro.

Depois que todos os 16 termos foram atribuídos alcançou-se uma perfeita concordância. Isto se deve a um efeito de 'saturação'. Há somente um número determinado de termos que se aplicariam de modo plausível a qualquer item, pelo menos se esses termos forem extraídos de um vocabulário controlado. Se forem atribuídos termos em número suficiente, acabar-se-á por alcançar uma elevada coerência. Observe-se, contudo, que a coerência baixa entre o nível de dois termos e o nível de dezesseis termos. Por exemplo, depois de cinco termos, o PC é 5/6 (0,83), depois de dez termos é de 6/14 (0,43), e assim sucessivamente.

A relação apresentada na figura 26 parece, portanto, plausível, embora não haja sido confirmada experimentalmente. Pelo menos a forma da curva é plausível, se se levam em consideração os resultados alcançados por muitos indexadores. No caso de poucos indexadores, naturalmente, o declínio da coerência seria provavelmente menos suave (por exemplo, haveria maior coerência com quatro termos do que com três).

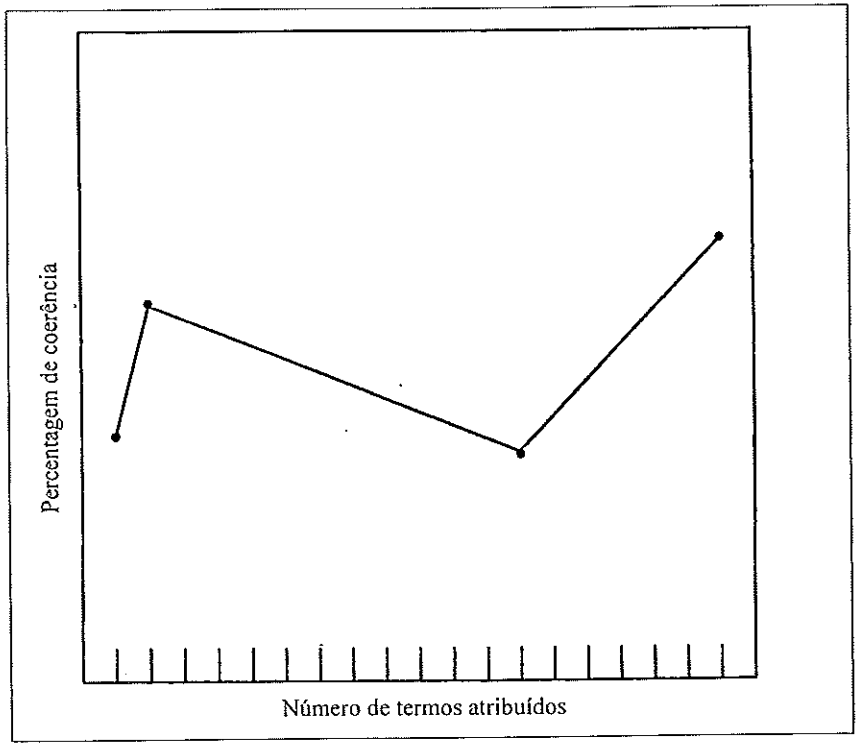


FIGURA 26
Relação entre coerência e quantidade de termos atribuídos

Harris et al. (1966) relatam resultados que diferem um pouco dos formulados hipoteticamente na figura 26. A coerência foi maior depois de 10 termos do que depois de cinco, mas declinou nos níveis de 20 e 30, voltando a aumentar quando foram atribuídos 40 termos. Afirmam que encontraram poucos indícios de algum efeito de saturação, mas seus indexadores utilizavam palavras-chave não-controladas e não as selecionavam de um conjunto limitado de termos controlados. Fried e Prevel (1966) descobriram um declínio da coerência com a quantidade de termos atribuídos, mas Leonard (1975) encontrou indicações inconclusivas sobre este ponto — verdadeiro para uma base de dados, mas não para outra.

Num estudo sobre coerência em bases de dados agrícolas, Reich e Biever (1991) encontraram prova do efeito da exaustividade sobre a coerência: numa amostra de artigos indexados com uma média de 8–9 termos de um tesauro, a coerência foi de 24%; chegou a 45% numa amostra que possuía uma média de 5–6 termos do tesauro.

a	*b*
A	B
B	A
C	D
D	E
E	K
F	C
G	G
H	L
I	M
J	N
K	O
L	P
M	F
N	H
O	I
P	J

FIGURA 27
Efeito da quantidade de termos atribuídos sobre a coerência do indexador (dois indexadores)

O segundo fator que influi sobre a coerência (figura 25) é o tipo de vocabulário utilizado na indexação. Uma das principais vantagens proclamadas para se adotar um vocabulário controlado é que ele melhorará a coerência na representação do conteúdo temático. Entretanto, a relação entre controle de vocabulário e coerência do indexador não é tão imediata quanto pareceria à primeira vista. Suponhamos que eu reúna alguns artigos de medicina e peça a um grupo de estudantes de nível médio que os indexem. Primeiro, exijo que façam a indexação extraindo palavras e frases dos próprios documentos. Eu pressuporia que haveria neste caso um razoável nível de coerência. Provavelmente, os estudantes se comportarão mais ou menos da mesma forma que um computador se comportaria ao realizar essa tarefa: procurarão palavras ou frases que ocorrem freqüentemente e/ou aparecem no título ou em outros lugares de destaque.

Numa segunda etapa desse exercício, peço aos estudantes que traduzam a indexação que fizeram com termos livres para termos selecionados no *Medical subject headings* (*MeSH*) da National Library of Medicine. Quase com certeza haverá uma queda da coerência. Nesta situação, o vocabulário controlado terá um efeito contrário. Isso se dá porque as expressões textuais selecionadas nem sempre serão idênticas aos termos controlados. Os estudantes terão dificuldade em selecionar os termos controlados apropriados porque carecem de conhecimentos suficientes de medicina e de sua terminologia e porque alguns dos termos controlados terão adquirido um 'significado' especial (indicado em nota explicativa) atribuído pelos compiladores do vocabulário. Um vocabulário controlado deve melhorar a coerência da indexação a longo prazo, mas somente pode

ser aplicado de modo coerente por indexadores experientes que dominem o conteúdo temático e estejam totalmente familiarizados com os termos.

Outra coisa para a qual é preciso atentar é que um vocabulário controlado deve melhorar a coerência da indexação em relação a um grupo de documentos, mas é bem possível que a diminua no nível de um único documento. Quer dizer, a terminologia adotada num artigo reveste-se de uma coerência interna — o autor costuma não empregar uma variedade de termos para descrever o mesmo tópico, pelo menos em artigos de natureza técnica ou especializada. É bastante possível, contudo, que haja divergência entre dois indexadores a respeito de qual o termo controlado a ser adotado para representar esse tópico. Por outro lado, autores diferentes empregam terminologias diferentes, e, desse modo, o vocabulário controlado, ao reduzir o leque de opções, exerce um efeito benéfico sobre a coerência da indexação quando se trata de um grupo grande de documentos.

Se a coerência interindexadores é baixa quando duas pessoas indexam itens que empregam o mesmo vocabulário, será natural, obviamente, que haja coerência ainda menor quando os mesmos itens são indexados em diferentes bases de dados porque variações nos vocabulários utilizados acrescentam outra dimensão ao problema. Qin (2000), por exemplo, reuniu um grupo de artigos sobre resistência a antibióticos e comparou a indexação deles no MEDLINE com a indexação baseada em citações feita no *Science Citation Index* (*Key Words Plus*). Naturalmente, a similaridade foi baixa, mesmo quando a 'similaridade parcial' era o critério adotado, embora as três palavras-chave usadas como termos de indexação que ocorreram com mais freqüência fossem conceitualmente equivalentes aos dois cabeçalhos de assuntos do MEDLINE de maior freqüência.

Convém salientar, de passagem, que não é tão simples quanto pareceria à primeira vista fazer uma comparação entre indexação com termos livres e indexação com termos controlados. Um termo controlado é ou não é atribuído. Na indexação com termos livres, entretanto, defrontamo-nos com o problema de ter que decidir se duas expressões são ou não são idênticas. Por exemplo, considera-se 'corrente elétrica' como igual a 'corrente eléctrica', ou como se avalia uma situação em que um indexador escolhe o termo 'literatura francesa medieval' e um outro utiliza 'literatura medieval' e 'literatura francesa'? Isso, é claro, nos faz remontar à distinção entre análise conceitual e tradução. Mais adiante se mencionará o efeito dessas duas etapas sobre a coerência.

Fugmann (1985) levanta uma questão muito interessante relativa à coerência. Salienta que, enquanto os estudos sobre coerência se concentram na seleção de termos para determinado documento, a pessoa que busca informação está mais preocupada com a coerência *entre* os documentos. Isso implica que talvez seja útil um tipo diferente de análise de coerência, que mensure a extensão com que o mesmo tópico é indexado coerentemente numa base de dados.

O terceiro fator identificado na figura 25 corresponde ao tamanho e à especificidade do vocabulário. Quanto maior o vocabulário, maior será a probabilidade de ser específico, e quanto maior for sua especificidade, mais difícil será utilizá-lo de modo coerente (Tinker, 1966, 1968). Por exemplo, há mais probabilidade de dois indexadores concordarem que um documento trata de corrosão do que de concordarem quanto ao tipo de corrosão que é estudado. Quanto mais sutis forem os matizes de significado que um vocabulário possa expressar, mais difícil será alcançar-se coerência. Na minha avaliação do MEDLARS (Lancaster, 1968a), incluí um breve estudo sobre coerência. Descobri que a coerência na atribuição de cabeçalhos de assuntos (*MeSH*) era de 46,1% quando os resultados de três indexadores eram divididos proporcionalmente entre um total de 16 artigos. Quando eram também utilizados subcabeçalhos, a coerência, no entanto, caía para 34,4%. Em estudo anterior verificou-se que os indicadores de função causavam efeito ainda mais drástico na redução da coerência da indexação (Lancaster, 1964), resultado esse que foi confirmado por Sinnett (1964) e Mullison et al. (1969).

Em seu estudo sobre coerência da indexação em bases de dados agrícolas, Reich e Biever (1991) concluem que "A coerência [...] parece ser mais difícil de alcançar à medida que aumenta a especificidade do vocabulário".

Slamecka e Jacoby (1963) fazem uma distinção entre vocabulários 'prescritivos' e 'sugestivos'. Estes oferecem ao indexador certa margem na escolha de termos, enquanto os primeiros praticamente não lhe deixam qualquer opção. Com base em alguns experimentos com vocabulários de diferentes tipos (cabeçalhos de assuntos, tesauro, esquema de classificação), concluíram que:

> A coerência interindexadores melhora significativamente com a utilização de instrumentos de indexação prescritivos que contenham um mínimo de relações semânticas variáveis entre os termos. O emprego de instrumentos de indexação que ampliem a liberdade semântica do indexador, no que concerne à escolha de termos, é prejudicial à confiabilidade da indexação. A qualidade da indexação tem muito a ganhar com vocabulários que formalizem as relações de modo a prescrever uniforme e invariavelmente a escolha dos termos de indexação (p. 30).

Assinale-se que eles parecem considerar coerência e qualidade como mais ou menos equivalentes. Este aspecto será estudado no capítulo seguinte.

É natural que os vocabulários prescritivos resultem em maior coerência. De fato, parece provável que se alcance o máximo de coerência com a atribuição de termos pré-impressos num formulário de indexação (como é o caso das 'etiquetas' da National Library of Medicine) que lembrem ao indexador que *devem* ser utilizados sempre que forem aplicáveis. Leonard (1975) apresentou algumas indicações que corroboram isso, do mesmo modo que Funk et al. (1983).

Leininger (2000), baseando-se em 60 itens acidentalmente indexados em duplicata na base de dados PsycINFO, verificou 66% de coerência na atribuição de etiquetas, enquanto a coerência total (considerados todos os termos) foi de apenas 55%. O resultado mais surpreendente foi que só houve 44% de coerência na atribuição de códigos genéricos de classificação. Com só 22 classes e 135 subclasses, e uma média de apenas 1,09 atribuições por registro (a maioria dos

registros é classificada apenas numa única classe e poucos num máximo de duas), seria natural que houvesse maior coerência. A explicação mais provável é que muitos artigos de psicologia parecem igualmente relevantes para duas ou mais categorias: embora indexadores diferentes concordem em qual de duas ou três categorias classificar um documento, haveria muito menor concordância quanto à única 'melhor' categoria. De novo as corridas de galgos e cavalos.

O quarto fator identificado na figura 25 corresponde à natureza do conteúdo temático do documento e, mais particularmente, sua terminologia. É de se supor que ocorra maior coerência na indexação de tópicos mais concretos (por exemplo, objetos físicos, pessoas designadas pelo nome), e que ela declinará à medida que se lidar cada vez mais com abstrações. Tibbo (1994) salienta que os autores da área de humanidades tendem a ser imprecisos em sua terminologia, preferindo textos 'densos' ao invés de legíveis. Entretanto, Zunde e Dexter (1969a) não verificaram aumento da coerência com a 'facilidade de leitura do documento'. Certos materiais podem suscitar problemas especiais no que tange à coerência da indexação. No caso de obras de criação, como livros de ficção, filmes de longa-metragem e alguns tipos de fotografias, é provável que haja um nível excepcionalmente elevado de desacordo em relação àquilo de que trata a obra e como indexá-la. Diferentes grupos de pessoas terão interesses bem distintos por esses materiais. Por exemplo, estudiosos das artes e do cinema talvez queiram uma indexação que seja bastante diferente daquela desejada pelo público em geral. Markey (1984) e Enser (1995) apresentam indícios que sugerem que a indexação de imagens pode produzir níveis de coerência excepcionalmente baixos.

O quinto fator tem a ver com os indexadores como indivíduos. É quase certo que dois indexadores com formação bastante similar (educação, experiência, interesses) tenham mais probabilidade de concordar com o que deve ser indexado do que dois outros com formação muito diferente. Relacionados a isso estão o tipo e a duração do treinamento. Se todos os indexadores participam do mesmo programa rigoroso de treinamento, isso contribui para reduzir a importância da formação prévia como fator que influi na coerência. Também é importante o conhecimento do conteúdo temático com que se lida. Se dois indexadores tiverem quase o mesmo nível de conhecimento especializado, serão mais coerentes entre si do que se um deles for muito entendido na matéria e o outro tiver apenas um conhecimento superficial do conteúdo temático. Mais importante do que o conhecimento especializado em si mesmo seria, contudo, o conhecimento minucioso das necessidades e interesses dos usuários a serem servidos.

Jacoby e Slamecka (1962) encontraram maior coerência entre indexadores experientes do que entre iniciantes que trabalhavam com patentes; os experientes também usavam menor quantidade de termos. Leonard (1975) constatou que a coerência aumentava com a experiência dos indexadores, mas não achou correlação positiva entre coerência e formação educacional. Quer dizer, maior conhecimento do conteúdo temático (presumido a partir da formação educacional) não aumentava a coerência. Korotkin e Oliver (1964), em experimento com resumos de psicologia, não descobriram diferenças significativas na coerência entre dois grupos de indexadores, sendo que um deles dominava o conteúdo temático e o outro não. Neste caso, porém, o estudo ocorreu sob várias restrições artificiais que iriam influir no resultado: não foi usado vocabulário controlado, foram empregados resumos e não artigos completos, e os indexadores foram instruídos a atribuir exatamente três termos (nem mais, nem menos) a cada item.

Um estudo posterior, de Bertrand e Cellier (1995), também examinou o efeito da experiência do indexador. Incluía, porém, tantas variáveis que se torna difícil interpretar seus resultados.

Dados encontrados em Stubbs et al. (1999) ilustram o efeito que um indexador 'radical' (isto é, atípico) pode provocar nos escores de coerência.

Outro fator apontado na figura 25 refere-se aos instrumentos auxiliares utilizados pelo indexador. Se um grupo de indexadores compartilhar o mesmo conjunto de ferramentas de indexação (dicionários, glossários, manuais), haverá uma tendência de que estes instrumentos contribuam para melhorar a coerência no grupo. O mais importante seria algum tipo de *vocabulário de entradas*, elaborado pelo próprio centro de informação, que servisse para correlacionar os termos que ocorrem nos documentos com os termos controlados apropriados.

Finalmente, a extensão do item indexado influi na coerência: quanto menor o item, menor será a quantidade de termos que a ele se aplicarão de modo plausível. Não causa espécie que Harris et al. (1966) hajam verificado que a coerência era maior na indexação de questões (breves enunciados textuais) do que na indexação de artigos de periódicos. Rodgers (1961), Fried e Prevel (1966), Leonard (1975), e Horký (1983) também encontraram indícios de coerência declinante com a extensão do documento, enquanto Tell (1969) constatou que a coerência quando se indexava a partir do texto integral dos artigos era menor do que quando se indexava a partir dos títulos ou dos resumos.

Coerência na análise conceitual versus coerência na tradução

O tipo de estudo de coerência examinado neste capítulo empana a distinção, que se faz na indexação, entre as etapas de análise conceitual e de tradução. Preschel (1972), porém, procurou separar essas duas etapas, a fim de determinar se era mais provável os indexadores concordarem com sua análise conceitual do que com a tradução em termos de indexação. Os resultados de sua pesquisa indicaram que era muito mais provável que os indexadores concordassem com o que seria indexado (análise conceitual) do que como os conceitos seriam descritos (tradução). É importante, porém, reconhecer que, nesse estudo, os indexadores não usaram um vocabulário controlado, mas criaram seus próprios 'rótulos verbais' para os tópicos. Resultados bem diferentes seriam alcançados se a influência normalizadora de um vocabulário controlado houvesse estado presente.

As figuras 28–31 mostram exemplos de conjuntos de termos de indexação

atribuídos a artigos por dois indexadores diferentes. Em todos os casos o vocabulário adotado foi o *Thesaurus of ERIC descriptors*. Todos são exemplos reais de enfoques alternativos na indexação. A indexação foi feita, como dever de casa, por alunos da Graduate School of Library and Information Science da University of Illinois. Os exemplos foram selecionados de um conjunto maior reunido pelo autor ao longo de anos. Os alunos tinham a liberdade de escolher os artigos que quisessem indexar, e era uma obra de puro acaso mais de um estudante escolher o mesmo artigo. Eles são aqui transcritos porque exemplificam alguns dos problemas que ocorrem na busca da coerência entre indexadores.

Indexador A	Indexador B
Termos mais importantes	*Termos mais importantes*
Vítimas de crimes	Assistência (comportamento social)
Assistência (comportamento social)	Formação de impressões
Apatia	Participação
Comportamento de quem busca ajuda	Testemunhas
Termos menos importantes	*Termos menos importantes*
Crime	Prevenção de crimes
Cidadania	Envolvimento
Esquiva	Leis
	Comportamento social
	Percepção social

FIGURA 28
Dois enfoques diferentes na indexação de um artigo intitulado "Quando os circunstantes apenas observam"

A figura 28 é um exemplo extremo: somente um termo em comum entre 16 atribuídos. O artigo trata do fenômeno de pessoas que se recusam a intervir quando testemunham um crime. Observe-se como os dois indexadores encaram o artigo de diferentes perspectivas — *B* mais do ponto de vista social e legal, e *A* mais do ponto de vista psicológico.

O exemplo da figura 29 não é muito melhor. Quanto aos termos mais importantes, os indexadores concordam apenas em relação a um deles. O artigo trata de um programa, oferecido por biblioteca pública, para instruir pais de crianças em idade pré-escolar sobre literatura adequada a esse grupo etário. O indexador *B* vê isso como educação pré-escolar, embora sejam os pais e não os filhos que recebam instrução, enquanto *A* (provavelmente de modo mais correto) acha que é educação de adultos pais. O indexador *B*, embora estudante de biblioteconomia, não indica que o programa acontece numa biblioteca. O indexador *A*, por outro lado, não indica que o artigo refere-se a crianças muito pequenas. Note-se como os dois escolheram termos relacionados muito próximos: *interesses de leitura* versus *atitudes diante da leitura*, *gosto pela literatura* versus *crítica literária*, *materiais de leitura* versus *seleção de materiais de leitura*.

Indexador A	Indexador B
Termos mais importantes	*Termos mais importantes*
Literatura infantil	Literatura infantil
Serviço de extensão em bibliotecas	Educação pré-escolar
Educação de adultos	Aspirações paternas
Educação de pais	Crítica literária
Seleção de materiais de leitura	
Termos menos importantes	*Termos menos importantes*
Relação pai–aluno	Experiência anterior
Leitura recreativa	Educação da primeira infância
Gosto pela leitura	Materiais de leitura
Interesses de leitura	Crianças pequenas
Ficção	Atitudes diante da leitura
Fantasia	Literatura
Bibliotecas públicas	Responsabilidade paterna

FIGURA 29
Dois enfoques diferentes na indexação de um artigo intitulado "Um curso de literatura infantil para pais"

Isso exemplifica os problemas inerentes ao uso de um vocabulário controlado que contém muitos termos bastante afins ou parcialmente coincidentes, principalmente quando os indexadores não estão totalmente a par do alcance pretendido desses termos.

A figura 30 mostra maior coerência, uma vez que dois dos termos mais importantes coincidem. Apesar disso, ocorrem algumas diferenças de tradução. O indexador *A* expressa 'cursos pós-graduados de educação' mediante o emprego dos termos *faculdades de educação* e *ensino superior*, enquanto *B* seleciona *faculdades de educação* e *ensino de pós-graduação*. De igual modo, quando *B* emprega *atitudes dos docentes*, *A* adota *opiniões*, e quando *B* usa *relação professor–aluno*, *A* emprega *relação interprofissional* e *orientadores pedagógicos*.

É difícil acreditar nos resultados da indexação da figura 31. Não existe termo algum em comum entre os doze atribuídos. Mais uma vez demonstram-se aí claramente os problemas decorrentes do emprego de termos afins e/ou coincidentes: são usados cinco termos sobre 'leitura', mas todos diferem entre si. Neste caso, porém, a indexação de *A* é bastante medíocre: não menciona o nível educacional e o item é indexado de modo muito genérico sob *ensino audiovisual* quando, especificamente, trata de televisão. Quando o documento foi indexado ainda não havia no tesauro o termo *televisão com legenda fechada*.

Os oito estudantes anônimos, cujo trabalho é comparado nas figuras 28–31, não eram indexadores altamente experientes, embora fossem inteligentes e interessados e estivessem motivados. É bastante provável que indexadores de maior experiência, principalmente com maior traquejo na utilização desse tesauro, houvessem alcançado resultados mais coerentes. De qualquer modo, os exemplos servem para ilustrar alguns dos obstáculos a uma indexação coerente.

Indexador A	Indexador B
Termos mais importantes	*Termos mais importantes*
Orientadores	Orientadores
Ensino superior	Faculdades de educação
Opiniões	Ensino de pós-graduação
Faculdades de educação	Atitudes dos docentes
Termos menos importantes	*Termos menos importantes*
Desenvolvimento profissional	Relação aluno–professor
Orientadores pedagógicos	Professores de pós-graduação
Orientação profissional	Estudantes de pós-graduação
Relação interprofissional	

FIGURA 30
Dois enfoques diferentes na indexação de um artigo intitulado
"Orientação em cursos de pós-graduação em educação"

Indexador A	Indexador B
Termos mais importantes	*Termos mais importantes*
Ensino audiovisual	Legendas
Pesquisa sobre leitura	Professores de televisão
	Ensino elementar
Termos menos importantes	*Termos menos importantes*
Ensino não-tradicional	Programas de remediação
Estratégias de leitura	Currículo de televisão
Motivação do aluno	Aptidões de leitura
	Ensino de leitura

FIGURA 31
Dois enfoques diferentes na indexação de um artigo intitulado
"Televisão com legenda fechada: uma nova ferramenta para o ensino da leitura"

A figura 32 é outra trapalhada. Aí dois estudantes registraram palavras e expressões que representam sua análise conceitual de um artigo, antes de tentarem traduzi-la em termos controlados. A comparação é muito instrutiva. Exceto o fato de ambos os conjuntos de termos se referirem a romances sentimentais, parecem ter pouco em comum. A interpretação de *A* é 'tranqüila' e romântica, enquanto o mínimo que se pode dizer de *B* é que é grosseira. *A* inclui somente três termos negativos (conflito, dominância, ressentimento), enquanto *B* inclui muitos termos radicais. O fato de serem possíveis tais interpretações radicalmente diferentes do significado de um artigo depõe, talvez, a favor do emprego da indexação como instrumento auxiliar da psicoanálise.

Embora duas ou mais pessoas possam não concordar rigorosamente com os termos que serão atribuídos a um documento, este fenômeno não é privilégio

Indexador A	Indexador B
Ficção romântica – Romances sentimentais	Mulheres como leitoras de ficção contemporânea
Ficção romântica de mulheres	Romances sentimentais
Conflito entre homens e mulheres	Heroínas
Relações de amor macho/fêmea	Fantasia feminina
Autopercepção feminina	Masoquismo–estupro
Dominância do macho sobre as mulheres	Romances góticos
Romances românticos como válvula de escape do ressentimento feminino	Papéis sexuais–estereotipagem
	Psicoanálise
	Auto-imagem feminina
	Narrativa
	Esquizofrenia
	Histeria
	Papéis sociais

FIGURA 32
Diferenças na análise conceitual de um artigo intitulado
"O ato em extinção: um estudo dos romances sentimentais"

exclusivo da indexação. Saracevic et al. (1988) constataram que os termos empregados para um mesmo pedido por diferentes especialistas em buscas revelavam uma coincidência extraordinariamente reduzida.* Além disso, itens recuperados por diferentes especialistas em buscas apresentavam pouca coincidência e cada especialista costumava encontrar alguns itens relevantes não encontrados pelos outros.** Saracevic sugere a necessidade de buscas múltiplas, feitas por diferentes pessoas, para o mesmo pedido, cujos resultados sejam reunidos e postos numa ordem classificada: os itens recuperados pela maioria dos especialistas ficarão no topo dessa classificação e aqueles recuperados apenas por um especialista ficarão na parte inferior. Pela mesma razão, um método ideal de indexação envolveria um trabalho de equipe, alcançando-se consenso sobre cada documento como resultado de discussões entre um grupo de indexadores. Ainda que este método tenha sido possível em alguns poucos locais altamente especializados (como os sistemas especializados existentes dentro do U.S. Patent and Trademark Office), ele é excessivamente dispendioso para a maioria das aplicações. Brown et al. (1996), entre outros, propuseram um método 'democrático' de indexação de imagens, em que os usuários da base de dados de imagens contribuíam com termos.

* Fidel (1985) também verificou que experientes especialistas em buscas mostravam pouca concordância na seleção de termos a serem empregados em buscas complexas. Anteriormente, Lilley (1954) e Bates (1977) mostraram que usuários de catálogos em fichas também costumam não concordar muito quanto aos termos a serem utilizados na consulta a esses catálogos.

** Katzer et al. (1982) constataram que representações diferentes de documentos faziam com que fossem recuperados diferentes conjuntos, os quais apresentavam pouca duplicidade mesmo quando as representações eram muito similares.

Bates (1986) sugere que a indexação é "indeterminada e probabilística" e que isso é mais ou menos inevitável, estando "arraigado na natureza da mente humana". Ao invés de lamentar o fato de que talvez jamais seja provável alcançar um elevado nível de coerência na indexação, pelo menos quando nela estão envolvidos indexadores humanos, devemos concentrar atenção na compensação disso na etapa final do processo, ou seja, no momento da busca. A busca não deve basear-se na coincidência exata de termos, mas em métodos que ordenem os documentos segundo o grau com que coincidem com alguma forma de enunciado de busca. A pessoa que executa as buscas deve dispor de diversos instrumentos auxiliares que lhe permitam selecionar dentre uma variedade de métodos para geração de associações semânticas entre termos.

Embora muitos estudos sobre coerência hajam sido realizados ao longo dos anos, muito poucas pesquisas foram feitas sobre *por que* diferentes indexadores selecionam diferentes termos, o que é sabidamente um tipo mais difícil de investigação. Dois artigos correlatos, de David et al. (1995) e Bertrand-Gastaldy et al. (1995), versam sobre este problema, mas chegam a conclusões bastante nebulosas.

Indexação coerente não é necessariamente o mesmo que indexação de alta qualidade. A qualidade da indexação será examinada no próximo capítulo, onde também se faz uma comparação entre qualidade e coerência.

CAPÍTULO 6

Qualidade da indexação

A indexação não constitui um fim em si mesma. Define-se de modo muito pragmático a 'boa indexação' como a indexação que permite que se recuperem itens de uma base de dados durante buscas para as quais sejam respostas úteis, e que impede que sejam recuperados quando não sejam respostas úteis. Cooper (1978) vai um pouco além:

> Justifica-se a atribuição de um termo a um documento se a utilidade média associada a essa atribuição for positiva, e injustificada se for negativa (p. 110).

Ele usa aqui a palavra 'utilidade' mais ou menos como sinônimo de 'benefício'.

Conforme as relações esquematizadas na figura 1 dão a entender, diversos subsistemas interagem no controle do desempenho de um sistema de recuperação da informação. Outro modo de examinar isso é em termos de uma seqüência de eventos que regem o desempenho da busca. Isso é exemplificado na figura 33.

Na situação típica de um centro de informação, uma necessidade de informação desponta na mente de um usuário desse centro e ele vai conversar sobre ela com um especialista em informação. Podemos nos referir ao resultado desse diálogo como um *pedido* (isto é, o entendimento por parte do especialista daquilo que o usuário realmente precisa). Com base nesse pedido, o especialista em informação prepara uma estratégia de busca, valendo-se para isso de termos de indexação, palavras do texto ou uma combinação de ambos. A estratégia de busca é então confrontada com a base de dados (é claro que, em muitos casos, a estratégia de busca e o cotejo com a base de dados estarão entrelaçados, pois a estratégia será desenvolvida interativamente em linha). Como resultado da busca certos itens são recuperados. Estes são peneirados pelo especialista em informação, a fim de eliminar todo item que lhe pareça evidentemente irrelevante, sendo entregue ao usuário um conjunto final de documentos ou referências.

O diagrama, naturalmente, representa buscas 'delegadas', ou seja, aquelas em que os clientes solicitam a um especialista em informação que localize para eles certas informações. Embora isso fosse a norma há uns vinte anos, cada vez mais deixa de ser assim, pois é crescente o número de pessoas que realizam suas próprias buscas em bases de dados acessíveis em linha, principalmente naquelas fontes acessíveis na Rede.

Com exceção do primeiro e último passos, porém, o diagrama ainda representa os fatores importantes que afetam o desempenho de uma busca temática numa base de dados. No caso de buscas não-delegadas, a necessidade de infor-

mação é diretamente convertida numa estratégia de busca num terminal sem passar pela etapa intermediária do 'pedido'.

Vê-se claramente, no diagrama, que muitos fatores influem na qualidade da busca, medida, por exemplo, pela revocação e precisão. Antes de mais nada, o especialista em informação precisa entender o que é que o usuário realmente precisa. Se o pedido for uma representação imperfeita da necessidade de informação, passa a ser quase irrelevante que todos os demais elementos — vocabulário, estratégia de busca, indexação, etc. — sejam satisfatórios.

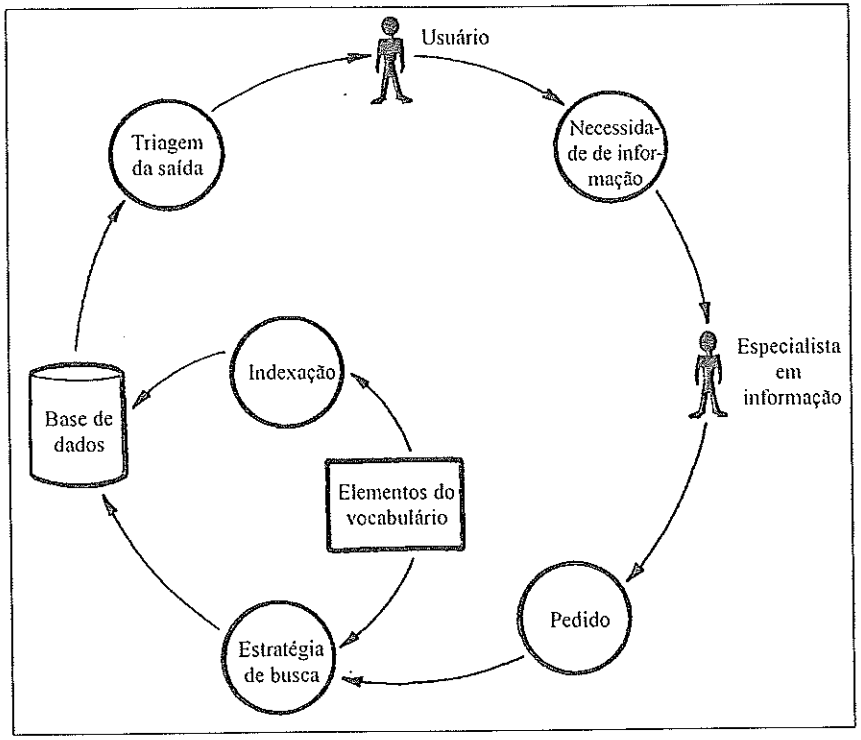


FIGURA 33
Fatores que influem nos resultados de uma busca numa base de dados

Admitindo-se que o pedido se aproxime razoavelmente da necessidade de informação, o fator seguinte a influir no desempenho será a qualidade da estratégia de busca. As principais influências a este respeito são experiência, inteligência e criatividade do especialista que faz a busca. O vocabulário da base de dados, contudo, também é essencial. Se for adotado um vocabulário controlado, não se poderá realizar uma busca que seja mais específica do que o vocabulário permite, embora se possa alcançar especificidade adicional com o emprego de palavras do texto. Infelizmente, é difícil imaginar todos os termos necessários à consecução de uma busca completa. O problema em todas as buscas é tentar

manter o equilíbrio entre revocação e precisão. O que se precisa comumente é obter o máximo de revocação, porém mantendo um nível aceitável de precisão.

Quando a estratégia de busca é cotejada com a base de dados, a qualidade da própria base torna-se, evidentemente, um dos fatores principais a influir no desempenho. É neste ponto, obviamente, que a qualidade da indexação se torna fundamental. Os elementos do vocabulário também influem na indexação, pois o indexador não pode lançar mão de termos que não existam no vocabulário.

A eficácia de uma 'triagem' do resultado, caso se efetue esta operação, dependerá fundamentalmente de dois fatores:

1. Em que medida o especialista que faz a busca entende aquilo de que o usuário realmente precisa.
2. Em que medida as representações de documentos armazenadas na base de dados indicam de que tratam os documentos.

Não convém fazer aqui uma análise minuciosa de todos os fatores que influem no desempenho de um sistema de recuperação, conforme esquematizado na figura 33, mas apenas examinar os fatores atribuíveis à indexação. Uma 'falha' de indexação pode ocorrer na fase de análise conceitual ou na de tradução.

As falhas de análise conceitual seriam de dois tipos:

1. Deixar de reconhecer um tópico que se revista de interesse potencial para o grupo usuário atendido.
2. Interpretar erroneamente de que trata realmente um aspecto do documento, acarretando a atribuição de um termo (ou termos) inadequado.

As falhas de tradução também seriam de dois tipos:

1. Deixar de usar o termo mais específico disponível para representar um assunto.
2. Empregar um termo que seja inadequado para o conteúdo temático devido à falta de conhecimento especializado ou por causa de desatenção.

Na prática, naturalmente, o avaliador de um sistema de informação não pode traçar algumas dessas distinções. Por exemplo, se o termo *X* for atribuído a um item quando não deveria sê-lo, não há como saber se o indexador interpretou equivocadamente qual seria o assunto do documento, se não entendeu realmente o significado ou alcance de *X*, ou se simplesmente atribuiu esse termo por descuido.

Se um indexador deixar de atribuir *X* quando este termo deveria ser atribuído, é óbvio que ocorrerão falhas na revocação. Se, por outro lado, for atribuído *Y* quando *X* é que deveria sê-lo, ocorrerão falhas tanto na revocação quanto na precisão. Quer dizer, o item não será recuperado quando de buscas de *X*, embora devesse sê-lo, e será recuperado em buscas de *Y*, quando não deveria sê-lo.

O descuido que leva à omissão de um termo que deveria ser atribuído ao documento pode ter profundo efeito nos resultados de uma busca, mesmo quando o termo omitido aparentemente não é importante. A figura 34 apresenta um exemplo simples disso, baseado num dos inúmeros que foram revelados durante a avaliação do MEDLARS (Lancaster, 1968a). O artigo trata do efeito sobre o de-

senvolvimento do córtex cerebral de nascimento ocorrido em situação de escuridão e permanente privação da luz. O indexador contempla todos os aspectos principais, menos o relativo ao desenvolvimento. Esta simples omissão será de grande importância. Neste caso, o artigo é considerado altamente relevante para um pedido de informação sobre fatores que influem no desenvolvimento do sistema nervoso central. O especialista em buscas somente usaria o termo 'desenvolvimento' para ter acesso a este tópico, pois seria irreal supor que pudesse prever que fatores seriam esses, e assim este artigo importante não seria recuperado.

No estudo sobre o MEDLARS, foram observados alguns exemplos de indexadores que empregaram termos incorretos, porém um número bem maior de casos de omissão de termos importantes por parte dos indexadores. Esta é provavelmente uma situação comum em outros serviços de informação.

Artigo	**Busca**
Tópico	*Pedido*
Efeito da privação da visão no desenvolvimento do córtex visual em camundongos	Fatores que influem no desenvolvimento, regeneração e degeneração do sistema nervoso central
Indexação	*Estratégia*
PRIVAÇÃO SENSORIAL ESCURIDÃO CÓRTEX CEREBRAL VISÃO CAMUNDONGOS	SISTEMA NERVOSO CENTRAL (hierarquia completa) *e* (DESENVOLVIMENTO *ou* REGENERAÇÃO *ou* DEGENERAÇÃO)

FIGURA 34

Exemplo da perda de um item importante por causa de mera omissão do indexador

Como reconhecer uma 'boa' indexação

A análise feita até agora neste capítulo implica que a qualidade da indexação somente pode ser aferida *ex post facto*, isto é, como resultado da experiência na operação de um sistema de recuperação e mais especificamente sua avaliação. Em grande parte isso é verdadeiro. Um conjunto de termos de indexação atribuídos a um documento não pode ser julgado 'correto' ou 'incorreto' em sentido absoluto. Ou seja, não existe nenhum conjunto 'melhor' de termos. Alegar que tal conjunto existe implica uma presciência de todos os pedidos que serão feitos à base de dados na qual o documento se acha representado.

Ocorrem, porém, realmente erros de indexação, e seria possível ao indexador experiente (ou 'revisor') descobrir pelo menos alguns desses erros antes da inclusão de um registro numa base de dados e assim impor certo controle de qualidade ao processo. Esse indexador identificaria os seguintes tipos de erros:

1. O indexador infringe a política, especialmente a relativa à exaustividade da indexação.

2. O indexador deixa de empregar os elementos do vocabulário da forma como devem ser utilizados (por exemplo, uma combinação incorreta de cabeçalho principal/subcabeçalho).
3. O indexador deixa de utilizar um termo no nível correto de especificidade. Na maioria dos casos isso significará que o termo selecionado não é o mais específico existente.
4. O indexador emprega um termo obviamente incorreto, talvez porque não possua conhecimento especializado (por exemplo, *combustíveis líquidos para foguetes* quando o documento trata mesmo é de combustíveis gasosos).
5. O indexador omite um termo importante.

Em primeiro lugar, o revisor comumente não despenderá, ao conferir a indexação de um item, tempo igual ao despendido pelo indexador. Talvez seja relativamente fácil reconhecer um termo incorreto, o qual provavelmente 'salta aos olhos' do indexador experiente, porém seria bastante difícil perceber a omissão de um termo importante, a menos que fosse muito óbvio (por exemplo, quando o termo aparece no título).

É possível testar o trabalho dos indexadores de uma maneira mais rigorosa do que simplesmente passando os olhos pelos termos atribuídos, que é o máximo que se pode esperar de uma operação rotineira de checagem. O método mais evidente consiste em realizar uma simulação de uma avaliação real. Consegue-se isso da seguinte forma:

1. Selecione um grupo de documentos dentre os que compõem o fluxo normal de entrada, antes que cheguem às mãos dos indexadores.
2. Para cada documento elabore, digamos, três questões para as quais o item seja considerado uma resposta importante. Uma das questões se basearia no tema central do documento enquanto as outras estariam centradas nos temas secundários, mas ainda assim importantes.
3. Faça com que experientes analistas de buscas elaborem estratégias de busca para cada uma dessas questões. É claro que esses analistas não devem ser as mesmas pessoas cuja indexação estará sendo examinada.
4. Faça com que os itens sejam indexados da forma rotineira.
5. Compare a indexação com as estratégias de busca, a fim de determinar se os itens relevantes são recuperáveis ou não com os termos atribuídos.

Como método para avaliar o desempenho de um grupo de indexadores, esse procedimento funcionará bastante bem se a amostra de documentos for suficientemente grande e se forem utilizadas as melhores estratégias de busca possíveis. Todo o teste seria realizado ao longo de uma série de semanas. Seria conveniente, naturalmente, que o mesmo conjunto de documentos fosse indexado várias vezes, uma vez por cada indexador, de modo que o desempenho dos indexadores fosse comparado sobre uma base comum. Isso, porém, nem sempre é possível devido à especialização de assuntos dentro do grupo.

Em grandes serviços de informação, que dependem do trabalho de muitos indexadores, especialmente quando a indexação é descentralizada, provavelmente será essencial implantar alguma forma de controle de qualidade. Se o volume de documentos indexados for muito grande, talvez seja economicamente inviável verificar todos os registros antes que dêem entrada na base de dados, e assim seria necessária alguma forma de amostragem. Seria possível, mas não suficiente, fazer uma amostragem completamente aleatória dos registros, principalmente se o índice de erros for provavelmente baixo. Isso exige um processo automático de 'marcar' os registros para que sejam inspecionados por especialistas, com base no fato de que tais registros parecem 'suspeitos'.

Todeschini desenvolveu um método engenhoso para identificar esses registros suspeitos (Todeschini e Farrel, 1989); Todeschini e Tolstenkov, 1990). Esse método vem sendo empregado na Agência Internacional de Energia Atômica, em Viena, para o controle de qualidade da base de dados INIS (Todeschini, 1997), e se tornou possível devido ao fato de os itens incluídos na base de dados serem indexados com descritores extraídos do tesauro INIS (uma média de aproximadamente 11 termos por item em 1990), além de serem classificados numa dentre 237 categorias genéricas de assuntos. Em essência, o sistema é capaz de identificar registros em que os descritores a eles atribuídos sejam atípicos dos descritores fortemente relacionados com a categoria onde foi anteriormente classificado. Se os descritores atribuídos a determinado documento, que houver sido colocado na categoria X, forem atípicos do 'perfil' do descritor anterior atribuído a X, esse registro será um bom candidato à revisão de controle de qualidade, pois a classificação ou a indexação pode estar errada.

Fatores que influem na qualidade da indexação

Lamentavelmente não foram muitas as pesquisas realizadas sobre os fatores que apresentam maior probabilidade de influir na qualidade da indexação. Na figura 35 apresenta-se uma tentativa de identificar esses fatores, mas ela se baseia mais no senso comum ou na intuição do que em provas concretas.

Os indexadores devem ter algum conhecimento do conteúdo temático tratado e entender sua terminologia, embora não precisem necessariamente ser especialistas no assunto. Na realidade, algumas instituições têm enfrentado problemas com indexadores que são 'especialistas' demais, pois sua tendência é interpretar o texto de modo excessivo e talvez extrapolar aquilo que o autor afirma (por exemplo, indexar uma aplicação possível que não esteja identificada especificamente no artigo) ou mesmo revelar preconceitos ao não indexar afirmações que relutam em aceitar (ver Intner, 1984, e Bell, 1991a, para comentários sobre viés e censura na indexação). A falta de conhecimento do assunto pode, contudo, levar à indexação excessiva. Incapaz de distinguir entre dois termos, o indexador talvez atribua ambos quando bastaria apenas um ou apenas um seria correto. Loukopoulos (1966) refere-se a isso como *indecisão* do indexador.

Fatores ligados ao indexador	*Fatores ligados ao documento*
Conhecimento do assunto	Conteúdo temático
Experiência	Complexidade
Concentração	Língua e linguagem
Capacidade de leitura e compreensão	Extensão
	Apresentação e sumarização
Fatores ligados ao vocabulário	*Fatores ligados ao 'processo'*
Especificidade/sintaxe	Tipo de indexação
Ambigüidade ou imprecisão	Regras e instruções
Qualidade do vocabulário de entradas	Produtividade exigida
Qualidade da estrutura	Exaustividade da indexação
Disponibilidade de instrumentos auxiliares afins	
	Fatores ambientais
	Calefação/refrigeração
	Iluminação
	Ruído

FIGURA 35

Fatores que podem afetar a qualidade da indexação

O autor declara-se reconhecido a Oliver et al. (1966) pela idéia que inspirou esta figura

Mai (2000) identifica cinco estádios no desenvolvimento de um indexador: principiante, principiante adiantado, competente, proficiente e especialista. Ele sustenta que somente o especialista tem capacidade para "indexar o mesmo documento com o emprego de diferentes métodos". Isso implicaria, por exemplo, que somente um especialista teria a capacidade de indexar o documento A para a clientela X e indexá-lo de modo diferente para a clientela Y. Ainda que isso soe aparentemente plausível, deve-se também admitir que é possível programar um computador para indexar o mesmo texto de diferentes formas (isto é, para diferentes clientelas) mediante a ligação de ocorrências de palavras/frases com diferentes conjuntos de termos de indexação.

É claro que um tipo particular de especialista é o próprio autor do documento. Já foram realizados alguns estudos sobre o autor como indexador. Por exemplo, Diodato (1981) estudou a coerência na seleção de termos entre três grupos: autores, indexadores e leitores de artigos de matemática. Ebinuma et al. (1983) traduziram as palavras-chave atribuídas pelo autor para os termos de um tesauro e os compararam com termos já atribuídos por indexadores experientes. A indexação oriunda do autor pareceu produzir melhor precisão porém menor revocação. Mulvany (1994) examina os prós e contras de os próprios autores indexarem seus livros.

Rasheed (1989) levou a cabo estudo similar, comparando termos atribuídos por autores de artigos de medicina com termos atribuídos por indexadores do MEDLARS. Ele constatou que os indexadores atribuíam muito mais termos e que os termos que eles empregavam eram mais específicos do que os empregados

pelos autores. Outros estudos trataram da indexação de livros como unidades independentes. Diodato e Gandt (1991) constataram que indexadores profissionais produziam índices que eram mais completos do que os índices feitos pelos próprios autores, embora as diferenças (por exemplo, em número de entradas por página de texto) não fossem tão grandes quanto seria de se esperar. Também se constatou que os autores apresentavam deficiências na redação de resumos de seus próprios artigos, aspecto a ser focalizado em próximo capítulo.

O conhecimento dos interesses dos usuários da base de dados é especialmente importante porque a 'boa' indexação deve ser talhada às necessidades de determinada comunidade, sempre que possível. Anos de experiência como indexador também são um fator que influi sobre a qualidade, da mesma forma que outras características, como a capacidade de a pessoa se concentrar, ler rapidamente e compreender prontamente. Finalmente, e talvez o mais importante de tudo, um bom indexador deve gostar do que faz. É improvável que se consiga obter uma boa indexação de alguém que detesta o que está fazendo.

Também intervêm nisso fatores ligados ao documento. Alguns assuntos são de mais difícil compreensão do que outros. Comumente, a teoria é muito mais difícil do que a prática, como ocorre nas diferenças entre mecânica aplicada e engenharia. Relacionado a isso, naturalmente, está o grau de 'correspondência' entre o conteúdo temático do documento e o conhecimento ou os interesses do indexador.

'Língua' pode ser interpretada de várias formas. Evidentemente, o indexador que não souber russo dificilmente poderá indexar artigos em russo de modo eficiente, a não ser que contenham resumos claros e completos na própria língua do indexador (o que não é usual). Outro aspecto concerne à clareza da linguagem do autor. Alguns autores expõem suas idéias ou descobertas de modo mais claro do que outros, tornando menos difícil o trabalho do indexador. Finalmente, existem alguns fatores ligados à apresentação que influirão sobre a maior ou menor facilidade que o indexador terá para descobrir de que trata o documento: o título é preciso ou enganador, existe um resumo ou algum outro tipo de sumarização que reflita integralmente o conteúdo do item?

É natural que os fatores ligados ao vocabulário também influam na qualidade da indexação. Quanto mais específico o vocabulário, mais minuciosos serão os matizes de significado que permite expressar; e quanto mais minuciosos os matizes de significado, mais difícil será estabelecer diferenças entre termos muito afins e empregar estes termos de modo coerente. Elementos sintáticos adicionais, como subcabeçalhos ou indicadores de função, aumentam a especificidade e complicam o trabalho de indexação.

Termos que sejam ambíguos ou imprecisos (que careçam de contexto adequado ou notas explicativas) são difíceis de interpretar e empregar corretamente, além do que o vocabulário deve contar com uma estrutura suficientemente completa (por exemplo, a estrutura TG/TE/TR do tesauro convencional) que guie o

indexador até o termo mais adequado para representar determinado tópico. As dimensões e a qualidade do vocabulário de entradas* também serão importantes, do mesmo modo que a disponibilidade de diversos instrumentos auxiliares afins, como dicionários ou glossários especializados.

Outros fatores que influem na qualidade têm a ver com o próprio processo de indexação. Alguns tipos de indexação, como a extração de palavras ou expressões do texto, não exigem muita concentração, esforço intelectual ou experiência, enquanto outros tipos, principalmente os que exigem o estabelecimento de relações conceituais precisas (mediante indicadores de função ou relacionais), encontram-se na extremidade oposta do leque de dificuldades. Em geral, é quase certo que os indexadores tenham desempenho mais eficaz quando recebem regras e instruções precisas do que quando trabalham em condições de completa liberdade. A produtividade exigida é outro fator importante. Se for exigido do indexador que dê conta de certo número de itens por dia, ele poderá sentir-se pressionado e isso levará a erros por descuido, especialmente se a instituição tiver uma expectativa excessiva de produção diária. Além disso, a indexação exaustiva demanda mais tempo do que a indexação seletiva.

Por fim, a indexação requer concentração, e condições ambientais desfavoráveis têm um efeito negativo sobre a exatidão dessa tarefa intelectual.

Outra maneira de analisar os fatores que influem na qualidade da indexação diz respeito às dificuldades que os indexadores defrontam. Oliver et al. (1966), em levantamento baseado em entrevistas, que abrangeu 61 indexadores, observaram que "tomar decisões sobre como melhor descrever o conteúdo dos documentos" era (o que não surpreende) o problema mencionado com mais freqüência. Infelizmente, este problema é geral, difuso e refratário a soluções fáceis. Outros problemas importantes mencionados foram 'entender material novo ou desconhecido' e falta de termos apropriados nos vocabulários controlados. Chu e O'Brien (1993) estudaram a etapa de análise conceitual da indexação, em pesquisa da qual participaram mais de uma centena de indexadores principiantes (estudantes), mas sua pesquisa baseou-se em somente três breves artigos, de modo que é difícil, a partir de seus dados, chegar a uma conclusão sólida.

A qualidade está relacionada à coerência?

Qualidade e coerência não são a mesma coisa: pode-se ser coerentemente ruim bem como coerentemente bom! Apesar disso, percebe-se intuitivamente que deve haver uma relação entre coerência e qualidade. Por exemplo, se três indexadores costumam concordar entre si, porém um quarto indexa de forma bastante diferente, a tendência da gente é acreditar no consenso.

---

* Um vocabulário de entradas é uma lista de termos não-preferidos, que ocorrem na literatura, que remetem para os termos preferidos apropriados mediante o emprego de remissivas do tipo ver ou usar. A importância disso é examinada alhures (por exemplo, em Lancaster, 1986).

Cooper (1969), em artigo polêmico, questiona o valor da coerência como indicador de qualidade. O aspecto por ele suscitado é exemplificado com referência à figura 36. Um centro de informação emprega quatro indexadores *A*–*D*. *B* e *C* são bastante coerentes entre si, porém *A* e *D* têm ambos suas idiossincrasias. No entanto, por essa ou aquela razão, a visão de mundo de *D* está mais próxima da dos usuários do centro, e os termos que atribui refletem melhor os interesses deles. Presume-se que sua indexação seja a melhor, pelo menos para essa clientela específica. Neste caso, então, os indexadores que são mais coerentes entre si não produzem o melhor trabalho, embora não sejam tão ruins quanto *A* cuja indexação se distancia ainda mais dos interesses dos usuários.

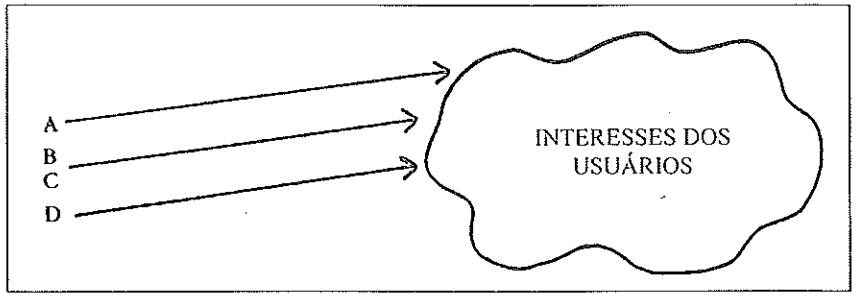


FIGURA 36
Coerência do indexador relacionada aos interesses dos usuários

Conquanto essa situação seja plausível, talvez não seja assim tão exagerada. É difícil compreender por que *B* e *C* seriam mais coerentes entre si, a menos que isso refletisse o fato de serem os indexadores mais experientes. Se o são, a lógica sugere que são esses dois os que deveriam ter mais conhecimento acerca dos usuários. São muito poucos os estudos que se relacionam de alguma forma com os argumentos de Cooper. No entanto, Diodato (1981) verificou, de fato, que a coerência entre autores de artigos de matemática e indexadores profissionais era maior do que a coerência entre autores e leitores dos artigos.

Leonard (1975) empreendeu o único esforço sério visando a estudar a relação entre qualidade e coerência na indexação. 'Qualidade' foi definida em termos de eficácia de recuperação — a capacidade de recuperar o que é desejado e de evitar o que não é desejado. Leonard trabalhou com duas coleções separadas de dados, que eram subconjuntos de estudos de avaliação anteriores. Essas coleções compreendiam documentos, pedidos, estratégias de busca e avaliações de relevância. Para cada pedido conheciam-se os itens que haviam sido julgados relevantes e quais os que não haviam sido considerados relevantes. Os conjuntos de termos atribuídos aos documentos pelos indexadores que participaram do estudo podiam assim ser comparados com estratégias de busca construídas anteriormente, permitindo ao pesquisador identificar se determinado documento seria ou não recuperado com determinada estratégia.

A comparação entre coerência e eficácia de recuperação mostrou-se mais difícil do que fora antecipado. Um problema importante se deve ao fato de que a 'eficácia' da indexação é normalmente associada ao trabalho de um único indexador, enquanto a coerência, por definição, é uma medida que se refere ao trabalho de dois ou mais indexadores (Leonard mediu a coerência do grupo bem como a coerência de par de indexadores). Leonard combinou os escores de 'eficácia' para dois (ou mais) indexadores e em seguida comparou este escore com a medida de coerência para estes indexadores. O escore de eficácia leva em conta a quantidade de documentos relevantes recuperados e a de documentos irrelevantes recuperados, e estes escores podem ser combinados determinando-se a média dos resultados para os dois indexadores ou agregando-os. Se se empregar o método de agregação, somente serão contados itens singulares, o que, com efeito, considera os dois indexadores como se fossem um indivíduo único.

Leonard observou uma relação positiva 'de moderada a forte' entre coerência e eficácia de recuperação, com uma 'relação positiva claramente definida' entre coerência e o coeficiente de revocação.

A utilidade dos estudos de coerência

A pesquisa realizada por Leonard (1975) sugere que de fato existe uma relação positiva entre coerência e qualidade da indexação, onde 'qualidade' refere-se à eficácia de recuperação. Mesmo que nenhuma relação houvesse sido descoberta, os estudos de coerência ainda teriam alguma utilidade. Hooper (1966) sugeriu várias aplicações, inclusive:

1. Na seleção ou treinamento de indexadores. A indexação feita por treinandos é comparada com algum padrão preestabelecido.
2. No controle permanente da qualidade das atividades de indexação.*
3. Para descobrir problemas na utilização de um vocabulário controlado; por exemplo, identificação de termos ou tipos de termos que sejam freqüentemente empregados de modo incoerente por causa de ambigüidades ou coincidências de sentido.
4. Para descobrir quaisquer problemas que possam existir relativos às regras de indexação.
5. Para determinar se a coerência é ou não menor no tratamento de certas áreas temáticas ou tipos de documentos.

Neste capítulo, aceitou-se que qualidade de indexação significa o mesmo que 'eficácia de recuperação' da indexação. Nem todos a definem desta forma. Rolling (1981), por exemplo, afirma que: "Pode-se definir qualidade de indexação como o grau de concordância entre os termos atribuídos pelo indexador e um

* Stubbs et al. (1999) examinam como os estudos sobre coerência interindexadores podem ser utilizados no monitoramento permanente da indexação numa instituição. Eles combinam cálculos de coerência com o emprego de 'cartas-controle' adotadas em engenharia industrial.

grupo de termos 'ideais' ou 'ótimos'." Em seguida, ressalta que a melhor maneira de alcançar o ideal é mediante alguma forma de consenso entre especialistas. O trabalho do indexador é comparado com o consenso, e ele seria 'penalizado' se não utilizasse termos sobre os quais os especialistas houvessem concordado, bem como se usasse termos sobre os quais não tivesse havido concordância. Rolling, que parece desconhecer o trabalho de Leonard, afirma que medidas de eficácia "não são praticáveis", enquanto os estudos de coerência "não são confiáveis". Ele defende estudos de qualidade, baseados no método do consenso, empregando-se os estudos de coerência apenas para pesquisar 'influências e tendências'. Mais no final deste capítulo encontra-se um exemplo da pontuação da indexação baseada nas sugestões de Rolling.

Vários outros pesquisadores procuraram avaliar a indexação fora do contexto do sistema de recuperação em que ela ocorre. Por exemplo, White e Griffith (1987) descrevem uma abordagem na qual são adotados métodos externos ao sistema de indexação que esteja sendo estudado, a fim de estabelecer um conjunto de documentos considerados 'similares em conteúdo'. Empregando conjuntos desse tipo (eles os denominam *aglomerados de documentos que servem de critério*) como base para avaliação, examinam três características dos termos de indexação atribuídos a itens do conjunto em determinada base de dados:

1. A extensão com que os termos unem itens afins. A medida óbvia disso é a quantidade de termos que foram aplicados a todos ou à maioria dos itens do conjunto. Os itens serão tidos como intimamente unidos se vários termos de assuntos houverem sido aplicados a todos eles.
2. A extensão com que os termos discriminam entre esses conjuntos na base de dados. A medida mais óbvia disso é a freqüência com que termos que se aplicam à maioria dos documentos do conjunto ocorrem na base de dados como um todo*. Termos muito comuns não são bons discriminadores. Por exemplo, no MEDLINE, o termo *humano* pode aplicar-se a cada item num conjunto, mas tem pouca utilidade para separar este conjunto de outros, uma vez que se aplica a inúmeros outros itens da base de dados. Por outro lado, termos que ocorrem muito raramente na base de dados como um todo serão úteis em buscas altamente específicas, porém terão pouca serventia na identificação de conjuntos um pouco maiores.
3. A extensão com que os termos discriminam minuciosamente entre documentos distintos. Aqui também a raridade é uma medida aplicável. Do mesmo modo é a exaustividade da indexação: um termo pode aplicar-se a todos os itens de um conjunto, mas não pode discriminar entre seus membros; quanto mais termos adicionais forem atribuídos a cada membro, mais diferenças individuais serão identificadas.

* Ajiferuke e Chu (1988) criticam o índice de discriminação adotado por White e Griffith porque não leva em consideração o tamanho da base de dados; propõem uma medida alternativa que leve isso em conta. Em artigo relacionado a esse (Chu e Ajiferuke, 1989), aplicam os critérios de avaliação de White/Griffith, com seu próprio índice de discriminação modificado, na avaliação da indexação em bases de dados de biblioteconomia.

Para examinar a qualidade dessa forma, deve-se primeiro estabelecer os conjuntos de teste, recuperar registros para os membros de cada conjunto de uma base de dados, e estudar as características dos termos atribuídos. White e Griffith empregaram essa técnica para comparar a indexação de seus conjuntos de teste em diferentes bases de dados. Comparar bases de dados dessa maneira é confirmar o pressuposto de que os itens do conjunto de teste são de fato similares em seu conteúdo. White e Griffith empregaram a co-citação como base para estabelecer seus conjuntos de teste, embora outros métodos, inclusive o acoplamento bibliográfico, também possam ser utilizados.

A utilidade desse trabalho é limitada pelo fato de que somente foram empregados aglomerados muito pequenos (na faixa de três a oito itens). Além disso, a validade do método como teste da indexação feita por seres humanos depende inteiramente de se estar disposto a aceitar um aglomerado de co-citações como sendo um padrão legítimo. Poder-se-ia apresentar um argumento convincente, segundo o qual faria mais sentido empregar indexadores especialistas como padrão para aferir a legitimidade do aglomerado de co-citações.

White e Griffith afirmam que o método é útil para um produtor de bases de dados aferir a qualidade da indexação, e apresentam exemplos de termos que talvez devessem ter sido utilizados pelos indexadores do MEDLINE ou acrescentados ao vocabulário controlado. Essas aferições de 'qualidade' podem, entretanto, ser feitas de modo mais simples: conjuntos de itens definidos por um termo ou termos determinados (por exemplo, 'supercondutores' ou 'supercondutividade', que ocorram como termos de indexação ou palavras do texto) são recuperados de diversas bases de dados e sua indexação é comparada sem o emprego da co-citação como padrão. Com efeito, este tipo de estudo também foi feito pelo mesmo grupo de pesquisadores (McCain et al., 1987). Para 11 pedidos formulados por especialistas nas ciências médicas comportamentais, foram feitas buscas comparadas nas bases MEDLINE, Excerpta Medica, PsycINFO, SCISEARCH e SOCIAL SCISEARCH. Nas três primeiras as buscas foram feitas com: a) termos controlados, e b) linguagem natural, e nas bases de citações foram feitas: a) empregando a linguagem natural dos títulos, e b) empregando citações de itens relevantes conhecidos como pontos de entrada. Embora o objetivo da pesquisa fosse estudar a qualidade da indexação do MEDLINE, pouco descobriu que se traduzisse em recomendações à National Library of Medicine quanto à prática da indexação, embora se fizessem recomendações sobre o alcance da indexação.

As conclusões mais importantes do estudo foram: 1) a incorporação de métodos de linguagem natural nas estratégias de busca resultou em melhoramentos significativos da revocação em comparação com o emprego somente de termos controlados, 2) a recuperação de citações deve ser considerada um complemento importante para a recuperação baseada em termos porque podem ser

encontrados itens relevantes adicionais com o emprego do método de citações, e 3) nenhuma base de dados pode sozinha fornecer uma cobertura completa de uma bibliografia multidisciplinar complexa.

A qualidade medida com o emprego de um padrão

Em estudo realizado para a National Library of Medicine (Lancaster et al., 1996), desenvolvi um método para avaliar a qualidade da indexação para o MEDLINE, seguindo a orientação proposta por Rolling (1981), que consistia em comparar o trabalho dos indexadores com um 'padrão', que seria um conjunto de termos estabelecido de comum acordo por indexadores altamente experientes. A figura 37 mostra o exemplo do padrão para um artigo e a figura 38 mostra os termos selecionados por dois indexadores diferentes para este mesmo artigo.

Escore	Cabeçalhos/subcabeçalhos
9 (6+3)	Auto-anticorpos/análise
26 (15+5+3+3)	Baço/*anormalidades/radiografia/radionuclídeo
6	Doença crônica
9 (6+3)	Doença de Hodgkin/cirurgia
6	Esplenectomia
6	Humano
6	Masculino
6	Meia-idade
9 (6+3)	Plaquetas/imunologia
6	Recidiva
6	Relato de caso
20 (15+5)	Tecnécio/*uso diagnóstico
15	*Tomografia computadorizada por raios X
23 (15+5+3)	Trombocitopenia/*imunologia/cirurgia
Total 153	

FIGURA 37

'Padrão' de indexação para um artigo médico, mostrando escores relativos à atribuição de vários tipos de termos

O padrão representa o consenso de um grupo de indexadores experientes sobre qual seria a indexação 'ideal' para esse item. Eles chegaram a 14 termos. Uns são cabeçalhos de assuntos, outros são etiquetas, e alguns dos cabeçalhos de assuntos recebem um ou mais de um subcabeçalho. Ademais, um cabeçalho de assunto ou uma combinação cabeçalho de assunto/subcabeçalho pode ser selecionado como 'mais importante'. Isto é, esses são os termos que os indexadores julgam mais importantes para o artigo e sob os quais o artigo aparecerá na versão impressa do *Index Medicus*. São identificados com um asterisco. Por exemplo, TOMOGRAFIA COMPUTADORIZADA POR RAIOS X foi selecionado como um termo mais importante, do mesmo modo que a combinação BAÇO/ANORMALIDADES. Note-se que o asterisco aplicado a um subcabeçalho é automaticamente transportado para o cabeçalho ao qual se acha ligado.

INDEXADOR A

Escore	Cabeçalhos/subcabeçalhos
15	*Baço
−13 (−7,−1,−4,−1)	Coristoma/complicações/*radiografia/radionuclídeo
8 (6,−1,+3)	Doença de Hodgkin/complicação/cirurgia
−4 (−3,−1)	Eritrócitos/radionuclídeo
−3	Espaço retroperitoneal
6	Esplenectomia
6	Humano
6	Masculino
6	Meia-idade
6	Recidiva
6	Relato de caso
7 (8,−1)	Tecnécio/uso diagnóstico
8	Tomografia computadorizada por raios X
10 (15,−4,−1)	Trombocitopenia/*etiologia/terapia
Total 64	

INDEXADOR B

Escore	Cabeçalhos/subcabeçalhos
15	*Baço
−15 (−7,−4,−4)	Coristoma/*radiografia/*radionuclídeo
6	Doença crônica
5 (6,−1)	Doença de Hodgkin/terapia
−4 (−3,−1)	Espaço retroperitoneal/radionuclídeo
4	*Esplenectomia
6	Humano
−4 (−3,−1)	Imunoglobulinas/uso terapêutico
−3	Indução de remissão
−3	Laparotomia
6	Masculino
6	Meia-idade
−4 (−3,−1)	Prednisolona/uso terapêutico
6	Relato de caso
15	*Tomografia computadorizada por raios X
6 (15,−4,−4,−1)	Trombocitopenia/*radiografia/*radionuclídeo/terapia
Total 29	

FIGURA 38

Escores de dois indexadores em comparação com o padrão da figura 37

O escore reflete a importância dos diversos termos e combinações de termos segundo o julgamento dos indexadores especialistas, a saber:

6 pontos por cabeçalho de assunto atribuído corretamente sem asterisco
6 pontos por etiqueta (à qual não se aplicam asteriscos)
3 pontos por subcabeçalho sem asterisco
15 pontos por cabeçalho de assunto sem asterisco
5 pontos por subcabeçalho sem asterisco.

O escore máximo possível para esse item é 153. Isto é, na hipótese muito impro-

vável de um indexador repetir exatamente o padrão, ser-lhe-ia atribuído o escore completo. Qualquer desvio do padrão — não atribuir um termo necessário, não usar o asterisco adequadamente, ou empregar um termo fora do padrão — resulta na perda de pontos. Note-se como os termos e as combinações de termos realmente importantes contribuem grandemente para o escore. O termo BAÇO leva três subcabeçalhos, um deles com asterisco. Baço faz 15 pontos porque recebe um asterisco do subcabeçalho com asterisco ANORMALIDADES, de modo que o escore total para esta combinação é de 15 para o cabeçalho principal com asterisco, cinco para o subcabeçalho com asterisco e três cada um para os outros dois cabeçalhos, num total de 26.

Esse item foi indexado duas vezes, uma pelo indexador A e uma pelo indexador B (figura 38). Pontuar o trabalho dos indexadores é um pouco mais complexo porque eles recebem uma pontuação positiva pela atribuição correta dos termos no padrão e uma pontuação negativa pela atribuição de termos que não sejam do padrão. Quando o indexador acerta exatamente o padrão para um termo, o escore para esse termo é transferido para o escore do indexador. Qualquer desvio resulta num escore reduzido ou, o que é pior, num escore negativo.

A pontuação completa é a seguinte:

Coincidência exata com o padrão: transportar o escore do padrão
–7 para cabeçalho com asterisco fora do padrão
–4 para subcabeçalho com asterisco fora do padrão
–3 para cabeçalho sem asterisco fora do padrão
–1 para subcabeçalho sem asterisco fora do padrão
4 para um cabeçalho com asterisco colocado pelo indexador, porém sem asterisco no padrão (ao contrário do 6 se o asterisco não fosse atribuído pelo indexador)
8 para um cabeçalho com asterisco no padrão, mas não colocado pelo indexador (ao contrário de 15 se o asterisco fosse atribuído corretamente)
– 1 para subcabeçalho com asterisco no padrão, mas que o indexador não atribuiu.

Muito embora isso pareça bastante complexo, não é bem assim porque, uma vez definido o método de pontuação, é possível escrever programas bem simples (e alguns já foram escritos) tanto para pontuar o padrão quanto para pontuar o trabalho dos indexadores em comparação com o padrão.

A aplicação de escores à indexação da National Library of Medicine é mais complexa do que o seria em muitas outras situações, devido ao emprego de subcabeçalhos e à distinção entre descritores mais e menos importantes, de modo que fica também mais difícil alcançar um acordo sobre quais devam ser os escores. Ainda que os escores numéricos verdadeiros usados nesses exemplos (reais) sejam considerados um tanto arbitrários, eles de fato refletem a enormidade percebida de vários tipos de erro indexado.

Se tiver havido acordo quanto aos escores, esse método de avaliação da

indexação é bastante discriminativo. Isto é, reflete claramente os desvios em relação ao padrão. Embora, nesse exemplo específico, nem o indexador A nem o indexador B tenham se saído muito bem, é evidente que A ficou mais perto do padrão do que B, e os escores refletem isso. B perdeu por ter deixado de fora por completo um termo considerado 'importante' pelo padrão e também porque introduziu vários termos externos ao padrão.

Como foi antes salientado neste capítulo, a qualidade da indexação é mais bem avaliada no contexto de uma avaliação completa do sistema de recuperação no qual são utilizados pedidos de usuários reais, como aconteceu no estudo sobre o MEDLARS (Lancaster, 1968a). Não obstante, a utilização do método do 'padrão-ouro' pode ser eficaz, especialmente na avaliação do progresso de indexadores em fase de treinamento e na comparação do trabalho de um grupo de indexadores com o de outro grupo.

Esse único exemplo ilustra também como a concordância quanto ao uso de etiquetas é muito mais fácil de alcançar do que a concordância quanto a outros termos, e que quanto mais refinada for a indexação (mediante o emprego de múltiplos subcabeçalhos e asteriscos) mais difícil fica alcançar acordo total.

Susanne Humphrey (1995), da National Library of Medicine, propôs um método de pontuação que usa escores de qualidade para medir a coerência da indexação. Nesse método, depois que os indexadores hajam sido pontuados em cotejo com o padrão, o trabalho pontuado que cada um executou num artigo torna-se o padrão em comparação com o qual cada um dos indexadores será avaliado, cada um por seu turno, no que tange à coerência. O emprego desse método pode ser ilustrado por meio de um exemplo simples, como o seguinte:

Indexador A		Indexador B	
A/a	6+3	A/a	6+3
B/c/d	6+3+3	C/*c	15+5
C/*c	15+5	D/d	6+3
		E	6
Total	41	Total	44

Se A for o padrão, o indexador B faz 29 pontos (os escores para os termos em que B concorda com A), de modo que a coerência é expressa como 29/41, ou 70,7. Se B for o padrão, o escore de A é de 29/44, ou 65,9. Quando as duas comparações (A com B, B com A) são combinadas, a média alcançada é de 68,3. Embora engenhoso, não fica totalmente claro qual o verdadeiro significado do escore. Basicamente, embora os escores de 'qualidade' hajam sido preservados, a qualidade não está sendo medida diretamente (pois o escore de nenhum dos indexadores é comparado com o padrão). Trata-se simplesmente de uma medida alternativa de coerência que, conforme foi sugerido no capítulo anterior, tem pelo menos o mérito de levar em conta a importância relativa dos termos. Isto é, se um indexador deixar de usar uma combinação de alta pontuação utilizada por outro indexador, isso reduzirá o escore de coerência entre eles muito mais do que o faria a falta de concordância quanto a um termo de baixo escore.

Capítulo 7

Resumos: tipos e funções

O resumo é uma representação sucinta, porém exata, do conteúdo de um documento. Endres-Niggemeyer (1998) adota definição semelhante: "Um texto, breve e coerente, que se destina a informar o usuário sobre os conhecimentos essenciais transmitidos por um documento".

É preciso distinguir entre as palavras *resumo* e *extrato*. Este é uma versão abreviada de um documento, feita mediante a extração de frases do próprio documento. Por exemplo, duas ou três frases da introdução seguidas de duas ou três frases das conclusões ou resumo do autor podem dar uma boa indicação daquilo de que trata um artigo de periódico. O verdadeiro resumo, ainda que inclua palavras que ocorram no documento, é um texto criado pelo resumidor e não uma transcrição direta do texto do autor. O termo 'sumarização' é hoje muito usado para designar qualquer processo que produza representações condensadas de textos e, assim, aplica-se tanto à redação de resumos quanto de extratos.

Os resumos podem ser caracterizados de inúmeras formas, inclusive segundo sua extensão. Na figura 3, por exemplo, apresentam-se dois resumos diferentes, um mais extenso do que o outro. Não há absolutamente razão alguma pela qual todos os resumos tenham aproximadamente a mesma extensão. Entre os fatores que influem na extensão de um resumo temos os seguintes:

1. *A extensão do item que está sendo resumido* (Craven, 1990, no entanto, não encontrou correlação entre a extensão do artigo e a extensão do resumo, porém ele trabalhou com uma área temática muito limitada);
2. *A complexidade do conteúdo temático*;
3. *A diversidade do conteúdo temático.* Por exemplo, um resumo preparado para os anais de um evento talvez precise ser bastante longo se os trabalhos apresentados abrangerem uma ampla gama de assuntos;
4. *A importância do item para a instituição que elabora o resumo.* Assim como ocorre com a exaustividade da indexação, um centro de informação industrial talvez precise redigir resumos mais longos dos relatórios da própria empresa do que de outros itens;
5. *A 'acessibilidade' do conteúdo temático.* Especialmente num serviço de resumos em forma de publicação, seria sensato fazer resumos mais completos de documentos menos acessíveis fisicamente (como relatórios de circulação limitada ou trabalhos apresentados em eventos) ou intelectualmente (por exemplo, redigidos em línguas pouco conhecidas).

6. *Custo.* Resumos longos não ficam necessariamente mais caros do que resumos curtos. De fato, talvez demore mais a redação de uma boa síntese de 200 palavras do que uma de 500. É óbvio, porém, que o custo de um serviço de resumos em formato impresso aumentaria de modo expressivo se a extensão média dos resumos aumentasse 50%, por exemplo. Isso teria reflexo sobre todos os custos, desde a composição do texto, até o papel e correio.
7. *Finalidade.* Um resumo que se destine essencialmente a proporcionar acesso a um documento com finalidade de recuperação precisa ser mais longo para que possa oferecer suficientes pontos de acesso.

Um resumo muito breve (por exemplo, que procure descrever um documento com uma única frase) é às vezes denominado *anotação*, termo que, no entanto, é bastante impreciso.*

Faz-se amiúde uma distinção entre *resumos indicativos* (às vezes denominados descritivos) e *resumos informativos.* Essa diferença é exemplificada nas figuras 39 e 40 que mostram dois tipos diferentes de resumos preparados para o item inicialmente apresentado na figura 3. O resumo indicativo simplesmente descreve (indica) de que trata o documento, enquanto o resumo informativo procura sintetizar a substância do documento, inclusive seus resultados.

> Foram feitas entrevistas telefônicas em 1985 com 655 norte-americanos selecionados por amostragem probabilística. Expressam-se opiniões sobre se: 1) a formação de um Estado palestino é essencial para a paz na região; 2) deve ser reduzida a ajuda norte-americana a Israel e ao Egito; 3) os EUA devem a) participar de uma conferência de paz que inclua a OLP, b) não favorecer nem Israel nem as nações árabes, c) manter relações amistosas com ambos. Os entrevistados indicaram se estavam suficientemente informados sobre os vários grupos nacionais da região.

Figura 39
Resumo indicativo

Isto é, o resumo indicativo mencionaria quais os tipos de resultados alcançados no estudo, enquanto o informativo faria uma síntese dos próprios resultados. Cremmins (1996) explica que os resumos indicativos contêm informações sobre a finalidade, alcance ou metodologia, mas não sobre os resultados, conclusões ou recomendações. Por outro lado, o resumo informativo inclui informações sobre objetivo, alcance e métodos, mas também deve conter resultados, conclusões ou recomendações. Para algumas finalidades, um bom resumo informativo serviria como um substituto razoável da leitura do documento.** É improvável

---

* Além de tudo, o campo da indexação e recuperação de vídeo costuma empregar 'anotação' ao invés de 'indexação', o que é imperdoavelmente enganoso.

** Isso não está isento de perigos. Por exemplo, Haynes et al. (1990) apresentam indícios que sugerem que os médicos às vezes tomam decisões sobre o tratamento dos pacientes baseados em leituras que não alcançam a totalidade do texto dos artigos médicos. Esse risco é agravado pelo fato de estudos recentes mostrarem que os resumos nas revistas médicas, mesmo as mais importantes, tendem a ser muito deficientes (ver capítulo 9).

que um resumo indicativo sirva como substituto dessa forma. Seu propósito principal seria indicar ao leitor do resumo se seria provável que viessem a querer ler o original. Por razões óbvias, os resumos informativos costumam ser mais longos do que os indicativos. Também são mais difíceis de redigir. Realmente, embora comumente seja possível redigir um resumo informativo de um estudo experimental, talvez isso seja quase impossível no caso de um estudo teórico ou um texto opinativo. Por isso, os resumos informativos ocorrem com mais freqüência nas ciências exatas e tecnologia do que nas ciências sociais ou humanidades.

Entrevistas telefônicas realizadas em 1985 com 655 norte-americanos, selecionados por amostragem probabilística, produziram estes resultados: a maioria (54–56%) acha que deve ser reduzida a ajuda norte-americana a Israel e ao Egito; a maioria (65%) é favorável à participação norte-americana numa conferência de paz que inclua a OLP; mais de 80% consideram importante que os EUA mantenham relações amistosas tanto com Israel quanto com os países árabes; 70% acreditam que os EUA não devem favorecer a nenhum dos lados; a maioria (55%) acha que a criação de um Estado palestino é essencial para a paz na região. Os israelenses são o grupo nacional mais conhecido e os sírios o grupo menos conhecido. A situação árabe-israelense só é superada pelo conflito na América Central entre os problemas internacionais mais sérios enfrentados pelos EUA.

FIGURA 40
Resumo informativo

Um mesmo resumo pode incorporar elementos indicativos e informativos (Cremmins refere-se a esse tipo de resumo como indicativo-informativo), dependendo dos interesses dos leitores que se têm em mira. Por exemplo, suponhamos um relatório sobre poluição atmosférica resumido numa publicação destinada a químicos. Grande parte do resumo, que trata dos aspectos ambientais, é meramente indicativa, mas uma parte dele será realmente informativa (por exemplo, apresentando resultados de análises feitas em amostras da atmosfera). Um mesmo serviço de resumos em formato de publicação pode conter tanto resumos indicativos quanto informativos. Geralmente, contudo, os resumos indicativos são mais comuns. Fedosyuk (1978) descreve procedimentos minuciosos para se distinguir entre resumos indicativos e informativos, valendo-se para isso de critérios lingüísticos e até mesmo apresentando um algoritmo com essa finalidade. Embora se trate de algo engenhoso, não se esclarece por que alguém precisaria de procedimentos formais para fazer essa distinção.

A expressão *inclinação para um assunto* é usada às vezes em relação aos resumos. Seu significado é que o resumo deve estar 'inclinado' para os interesses dos usuários que se têm em mira. Ou seja, na redação de resumos, bem como na indexação, a pergunta norteadora deve ser: "Por que nossos usuários provavelmente se interessarão por este item?" Os resumos preparados por uma instituição para serem usados internamente estarão sempre inclinados para as necessidades e interesses locais. A situação é um pouco mais complicada no caso de serviços de resumos em forma de publicação.

Faz-se diferença entre serviços *orientados para uma disciplina* e os *orientados para uma missão*. Os primeiros buscam atender às necessidades de uma disciplina (por exemplo, química, biologia, ciências sociais) enquanto os últimos procuram ir ao encontro das necessidades de determinada indústria ou grupo de indivíduos (por exemplo, resumos para a indústria da borracha ou resumos para enfermeiros). A inclinação para um assunto é mais relevante e viável no caso de serviços orientados para uma missão do que para os que se orientam para uma disciplina, porque os interesses dos usuários dos primeiros costumam ser mais homogêneos e especializados do que os interesses dos usuários dos últimos. Pelo menos um estudo mostrou que bem pouca inclinação para um assunto ocorre em serviços de resumos em formato impresso (Herner, 1959).

Outro tipo de resumo é o *resumo crítico*. Trata-se, com efeito, de uma 'recensão crítica condensada'. Aplicado a relatórios, artigos de periódicos e outros itens relativamente breves, o resumo crítico serve quase ao mesmo propósito de uma recensão crítica de livro. O resumo crítico é avaliador. O resumidor opina sobre a qualidade do trabalho do autor e pode até compará-lo com o de outros. Por exemplo, um resumo crítico do item mostrado na figura 3 mencionará as deficiências da metodologia utilizada — a maneira como se obteve a amostra da população, o tamanho da amostra, a maneira com as questões foram formuladas — ou comparará os resultados com os de pesquisas anteriores. Como os redatores devem ser especialistas de fato, os resumos críticos são bastante raros.

Duas publicações que anunciam a característica de incluir resumos críticos são *Mathematical Reviews* e *Applied Mechanics Reviews (AMR)*. A figura 41 mostra um resumo crítico real reproduzido da última dessas publicações. Note-se que o resumo é assinado e combina elementos descritivos e críticos. Uma análise da *AMR* revela, porém, que resumos verdadeiramente críticos sempre foram muito mais a exceção do que a regra, e hoje em dia eles não aparecem nessa revista, que somente se acha disponível em formato eletrônico em linha.

Atualmente, os resumos aparecem freqüentemente em periódicos científicos junto com os artigos a que se referem; são comumente redigidos pelos autores dos artigos. Em muitos casos esses resumos são reproduzidos pelos serviços de índices e resumos. Alguns periódicos incluem resumos em mais de uma língua. Por exemplo, muitos periódicos russos e japoneses incluem resumos em inglês.

Finalidade dos resumos

Poderíamos mencionar muitas e diferentes finalidades dos resumos. A mais importante, talvez, é que os resumos facilitam a seleção. Ou seja, ajudam o leitor a decidir se determinado item apresenta a possibilidade de satisfazer a seu interesse. Desse modo, poupam tempo ao leitor, evitando, por exemplo, que obtenha artigos que não teriam interesse para ele. Em alguns casos, também, um bom resumo informativo pode realmente substituir a leitura de um item que seja de interesse para o usuário. Os resumos são particularmente úteis para esclarecer

1989. Pao, Y. C., Dept. of Eng. Mech., Univ. of Nebr., Lincoln, Shy, D.S., et al., On relationship between bulk modulus and relative volume of lung during inflation-deflation maneuvers, p 136-142, *Journal of Biomechanical Engineering, Transactions of the ASME* v 104 n 2 (May 1982).

The paper presents an equation relating the bulk modulus of the lung to the relative volume during inflation and deflation. The average bulk modulus of the lung was obtained by injecting air via a 6-mm-i.d cannula in the main lobar bronchus. "Regional lobe" volume changes were measured by roentgen-videographically determined placement of 25 metal markers implanted in the excised lower lobes of three dogs. Whole lobe volumes at various transpulmonary pressures were measured by water displacement. Pressure and volume measurements were used to calculate bulk modulus ($K = \Delta V P / \Delta V$). The "most satisfactory least squares curve-fit" of bulk modulus (K) vs. relative volume (V / V_{max} was obtained with the equation $K = C/(1 - V/V_{max})^n$. Substituting for bulk modulus with the equation $K = VdP/dV$, and integrating enabled computer-generated pressure-volume plots. This equation provided a better pressure-volume curve-fit than previously obtained, especially at low values of pressure and volume. Also, as expected, the bulk modulus was smaller at low volume, but the rate of change of modulus was greater during deflation than during inflation.

The authors assumed, without giving sufficient justification, that the "regional lobe" (the area bounded by the 25 markers) included a higher density of airways than the rest of the lobe. Using this assumption, the authors claimed that the modulus and rate of change of modulus were different for parenchyma tissue and the airways during both inflation and deflation. No mention, however, was made of paired t-tests or any other statistical tests. In fact, if they had done a paired t-test, they would have discovered that none of these differences were significant, even at the 90 percent confidence level. Other source of errors which were not addressed include: the difference in the properties of excised lung and intact lung due to blood in the vessels, surrounding tissue,negative pressure, etc.; the effect of the markers on the pressure-volume relationship; the effect of strain rate on the modulus of lung tissue, which is a viscoelastic material; the time elapsed between regional volume measurement and whole volume measurements (this is important for viscoelastic material); the difference between the true regional ΔV; and the measured ΔV; and the differences between the mechanical properties of dog and human lung tissue.

Despite its limitations, the paper presents a step forward in the understanding of mechanical properties of the lung, and, thus, lung diseases. Therefore, it should be of benefit to researchers interested in respiratory mechanics and physiology.

D. S. Feldman, USA

FIGURA 41
Exemplo de um resumo crítico
Reproduzido de *Applied Mechanics Reviews*, 37, 1984, com permissão da editora

o conteúdo de documentos escritos em línguas que o leitor desconheça. Janes (1991) descobriu, o que não causou surpresa, que os resumos eram mais eficientes do que outras partes do registro, como títulos e termos de indexação, na avaliação da relevância de um item.

A impressão e distribuição de resumos é um meio eficaz para manter as pessoas informadas a respeito da bibliografia recentemente publicada em seus respectivos campos de interesse (isto é, proporcionando-lhes um serviço de alerta

ou *notificação corrente*). Conforme foi mencionado anteriormente, os resumos que acompanham artigos ou relatórios são úteis para o indexador na medida em que o ajudam a identificar, do modo mais rápido possível, o conteúdo temático dominante do documento. Borko e Bernier (1975) sugerem que os resumos podem substituir o texto integral nas atividades de indexação, porém esta é uma prática nem sempre conveniente.

Finalmente, os resumos desempenham atualmente importante papel nos sistemas de recuperação informatizados porque facilitam a identificação de itens pertinentes e proporcionam acesso a itens armazenados (nos sistemas em que o texto dos resumos é armazenado em formato que se presta à recuperação). Levando em conta tanto a revocação quanto a precisão, foi demonstrado que os processos automáticos de recuperação baseados em resumos eram mais eficazes do que aqueles baseados nos textos integrais dos documentos (Lam-Adesina e Jones, 2001), embora ainda faltem mais evidências acerca desse ponto.

Hartley e Benjamin (1998) alegam que os resumos cresceram de importância ao longo dos anos na medida em que cresceu a literatura científica:

> Na realidade, a natureza dos resumos alterou-se ao longo dos anos, na medida em que mais e mais artigos científicos passaram a competir entre si para atrair a atenção dos leitores. Hoje os leitores precisam compulsar e pesquisar mais do que o faziam no passado, e o resumo evolui continuamente como um portal de acesso à literatura científica (p. 451-452).

Na ciência, salientam eles, os resumos estão ficando mais extensos e mais orientados para os resultados.

Para certas finalidades, o *resumo estruturado* é preferível a um resumo em formato de texto narrativo. Um exemplo hipotético do 'gabarito' de um resumo estruturado é mostrado na figura 42. O conteúdo temático de que trata é irrigação. Neste caso, o resumidor é solicitado a procurar especificamente os itens listados. A elaboração do resumo consiste em colocar os 'valores' apropriados no gabarito. Quer dizer, indicam-se para cada artigo o tipo de irrigação, o tipo de solo, os produtos cultivados, as condições climáticas e a localização, sendo empregados códigos que representam os tipos de resultados obtidos. Este tipo de resumo é útil na compilação de manuais que sintetizam um grande número de estudos realizados em determinado campo. No entanto, só daria certo numa área temática em que os elementos essenciais permanecessem mais ou menos os mesmos entre os diferentes estudos. Zholkova (1975) descreve como se adotaria a análise de facetas para criar um resumo estruturado, mas não chega a convencer quanto à utilidade desse método.

Hartley et al. (1996) compararam resumos estruturados com resumos não-estruturados numa atividade de busca de informação. Observaram que os sujeitos de seu experimento podiam usar os resumos estruturados de modo mais eficaz (isto é, com maior rapidez e/ou menos erros) na localização de respostas a consultas ou na identificação de resumos que fossem pertinentes a determinado

tema. No entanto, a forma como usam o termo 'estruturado' é muito diferente da minha. Para eles, um resumo estruturado é simplesmente o que traz entretítulos (histórico, objetivo, métodos, resultados, conclusões) para facilitar a rápida visualização do texto (e do modo como é hoje usado em muitas revistas médicas), enquanto eu uso o termo para designar o resumo redigido em formato não-narrativo. O tipo de resumo estruturado da figura 42 poderia concebivelmente ser produzido com o uso de um programa de computador projetado para identificar e extrair do texto os valores apropriados (ver os comentários sobre o método de preenchimento de padrão para extração e sumarização de textos nos capítulos 14 e 15. Em alguns lugares, o tipo de resumo analisado por Hartley et al. foi simplesmente designado como 'resumo mais informativo' (Haynes et al., 1990; Haynes, 1993), e acredito ser esta uma melhor denominação. Este tipo de resumo estruturado será visto com mais vagar no próximo capítulo.

TIPO DE IRRIGAÇÃO	TIPO DE SOLO	PRODUTOS	CONDIÇÕES CLIMÁTICAS	LUGAR	RESULTADOS

FIGURA 42
Gabarito para um resumo estruturado

Um tipo totalmente diferente de resumo estruturado, em formato de diagrama, foi proposto por Broer (1971). Como mostra o exemplo inteiramente fictício da figura 43, o resumo parece um diagrama em bloco, ou fluxograma, em que blocos interconectados de palavras, com títulos padronizados, são usados para expressar a essência do artigo. Broer diz que essa forma de resumo é mais fácil de examinar e compreender, e mostra um resumo convencional para comparação (figura 43). É uma proposta curiosa, mas nunca se popularizou. Uma desvantagem, que é o espaço ocupado na página impressa, não existiria na visualização em linha, de modo que talvez a internet possa reacender o interesse por esse formato.

Bernier e Yerkey (1979) descreveram e exemplificaram o emprego de enunciados altamente condensados, cada um sintetizando o 'ponto' mais importante de uma publicação. Referem-se a esses enunciados genericamente como 'literaturas concisas' e à sua forma mais condensada como 'literaturas ultraconcisas'. Uma variedade é a conclusão ultraconcisa, um enunciado bastante sucinto sobre a conclusão mais importante alcançada por uma pesquisa. Por exemplo:

A lingüística teórica não teve qualquer impacto importante na ciência da informação

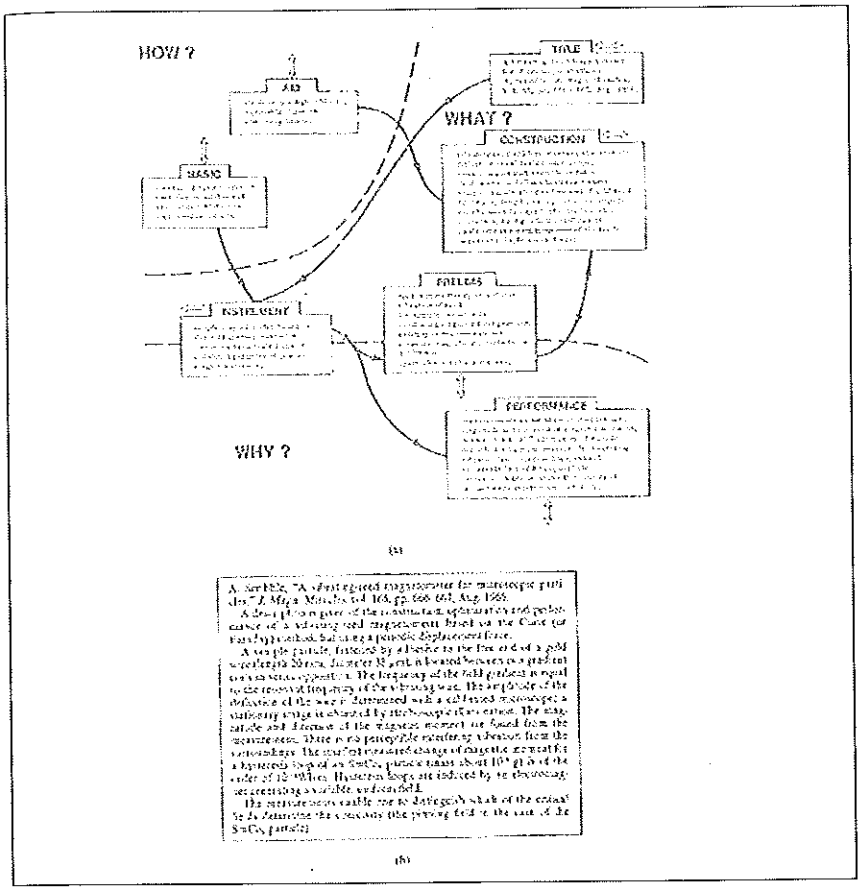


FIGURA 43
Resumo em 'diagrama de bloco' de um artigo hipotético junto com um resumo 'convencional' para comparação
Reproduzido com permissão de J.W. Broer, "Abstracts in block diagram form", *IEEE Transactions on Engineering Writing and Speech* (© 1971, Institute of Electrical and Electronics Engineers)

Este tipo de sumarização não é um resumo no sentido convencional; no entanto, as literaturas concisas certamente guardam uma relação com os resumos. Apresentam muitas aplicações potenciais. Por exemplo, seria possível produzir um manual que condensasse o que se conhece acerca de determinado fenômeno (por exemplo, uma doença) na forma de uma série de enunciados ultraconcisos, sendo cada um desses enunciados acoplado a uma referência bibliográfica que identificaria a fonte de onde foi extraído.

Resumos modulares

Em 1964, Herner and Company realizou um estudo para a National Science Foundation sobre a viabilidade de 'análises de conteúdo modulares' (Lancaster et al., 1965). Elas continham dois componentes: resumos modulares e entradas de índice modulares. Nas figuras 44–45 apresenta-se uma amostra disso.

Citation

Rosensweig, R. E., and Beecher, N. Theory for the ablation of fiberglas-reinforced phenolic resin. American Institute of Aeronautics and Astronautics Journal, vol. 1, No. 8, August 1963, pp. 1802-1809.

Annotation

A theoretical model is developed, for a charring and melting composite material, combining glassy ablation and the char layer-molten glass chemical reaction effects.

Indicative

The variables associated with the ablation of a typical resin-glass system are examined. These include glass ablation and plastic pyrolysis, flow in both the reacting and non-reacting parts of the melt, mass loss and heat absorption due to chemical reaction, mass injection effects, and coupling between the external pressure and the assumed chemical reaction. The mathematical development is traced and the approximations utilized are discussed. Parametric examinations are made.

Informative

Pyrolysis, melting, and chemical reaction are taken into account in this theory of the ablation of phenolic-fiberglas. It postulates a very thin, isothermal, surface reaction zone, where the char layer (carbon) formed during the pyrolysis of the organic binder reacts chemically with the molten silica. Other assumptions are conventional.
Calculations for typical IRBM re-entry conditions showed little temperature drop in the reaction zone, 6% maximum and usually less than 1%. Depth of the zone was three orders of magnitude less than the thermal thickness. The unreacting run-off in the melt zone ranged from 40-80% as a function of the possible reaction enthalpy level. However, more than 99% of the material reaching the reaction zone was affected. At the expected temperatures of 1400-2000°C, the theory assumed the reaction

$$SiO_2 + 3C \rightarrow SiC + 2CO$$

Earlier experiments had yielded the reaction kinetics. Significant effects, up to 25% increase, on the ablation rate appeared only at the lower reaction rates. Changing the reaction enthalpy by a factor of three changed the ablation rate by less than 10%. When compared with a peak re-entry ablation rate, the value given by this theory was reported to be 38% in defect.

Critical

This theory extends the classic work of Bethe and Adams (Avco-Everett Research Lab., Res. Rept. 38, Nov. 1958) on ablation of pure glasses. Thus it treats the problem as concerning carbon-contaminated glass rather than, as is more usual, a char-layer. In the only comparison given between the theory and experimental data, revealing 38% underprediction by the theory, a thorough error analysis was not included. Spalding (Aero. Quart., Aug. 1961, pp. 237-274) and Scala (General Electric Co. (MSVD), Rept. R59SD401, July, 1959, ARS Jnl., June, 1962, pp. 917-924) have treated similar problems.

FIGURA 44
Resumos modulares

Os resumos modulares destinavam-se a ser descrições completas de conteúdo de documentos correntes. Cada um possuía cinco partes: citação, anotação, resumo indicativo, resumo informativo e resumo crítico. O conjunto fora planejado de modo que um serviço de resumos podia processá-lo para adaptá-lo a seus próprios requisitos com o mínimo de esforço: qualquer resumo seria utilizado na íntegra, ou os módulos teriam o texto reorganizado para formar, por exemplo, um resumo parcialmente indicativo, parcialmente informativo, ou um resumo parcialmente informativo, parcialmente crítico.

A finalidade primordial dos resumos modulares era eliminar a duplicação e o desperdício de esforço intelectual envolvidos na elaboração, de forma independente, de resumos dos mesmos documentos por vários serviços, sem qualquer intenção de impingir resumos 'padronizados' a serviços cujas exigências variam notavelmente quanto à forma e à inclinação para um assunto. Tanto os resumos quanto as entradas de índice eram preparados por especialistas no assunto, e a intenção era de que eles conciliariam os requisitos de rapidez de publicação com a meticulosidade de resumos preparados por especialistas. Seu formato e tratamento padronizados também reduziriam o processamento repetitivo e acelerariam o fluxo de trabalho nos serviços de resumos beneficiários.

Physical and Mathematical Systems

Axisymmetric and Blunt Body Systems
Re-entry Bodies

Thermodynamics

Coupled Reactions
Carbon-Silica Reactions

Environment

Atmospheric Entry
Re-entry Conditions
Space Flight

Materials

Phenolics, Fiberglas Reinforced
Glass Fibers
Rocket and Missile Materials
Ablation Materials
Reinforced Plastics
Thermal (Re-entry) Shields
Phenolic Resin

Mass Transfer

Ablation, Analytical
Ablation, Charring
Ablation, Melting
Ablation of Glasses
Chemical Reaction Effects
Thermal Thickness
Reaction Zone
Reaction Thickness
Gasification Ratio

Means and Methods

Parametric Analysis

Authors

Rosensweig, R. E.
Beecher, N.

Affiliations

Massachusetts Institute of Technology
National Research Corporation

FIGURA 45
Entradas de índices modulares

As entradas de índices modulares sugeriam termos descritivos, extraídos de vocabulários de indexação representativos, que poderiam ser utilizados completos, com aperfeiçoamentos ou acréscimos, para indexar o resumo oriundo do pacote modular. Os vocabulários de indexação representativos, utilizados como fontes para as entradas do índice modular, seriam extraídos dos índices correntes ou de listas autorizadas dos serviços de resumos e indexação participantes, refletindo assim os estilos e políticas de indexação desses serviços.

Testou-se essa proposta no campo da transferência de calor, pois, sendo este assunto altamente interdisciplinar, revestia-se de interesse potencial para inúmeros serviços de resumos. Conjuntos de resumos/entradas de índice foram preparados e submetidos à apreciação de diversos serviços para que fossem processados rotineiramente. Esses serviços preencheram questionários de avaliação da proposta. A conclusão foi que era possível produzir uma análise de conteúdo, em forma modular, que seria adotada como entrada por vários serviços de resumos, mas que a maioria deles relutava em abrir mão de sua autonomia a fim de participar do tipo de centro referencial implícito no método modular.

SUMMARY

1. A method is described for the determination of strontium and barium in human bone by radioactivation analysis.
2. Results of analyses of 35 bone samples, from normal persons of both sexes and different ages, are given. The concentrations of barium and strontium were found to be of the order of 7 and 100 μg./g. of ashed tissue respectively.
3. No relationship between sex or disease of individuals with strontium and barium concentration was noted. The concentration of strontium in the age group 0-13 years was significantly lower than that in the group 19-74 years.
4. No significant difference was found in the concentrations of strontium and barium in the various bones of those individuals examined.
5. Results obtained in this survey are discussed and compared with those of other workers.

/00193/
/METHOD/DETERM/STRONTIUM/BONE/HUMANS/RADIOACTIVATION ANALYSIS/
/00193/
/NO RELAT BETW/STRONTIUM/HUMANS/AND/SEX/OR/DISEASE/
/00193/
/NO RELAT BETW/BARIUM/HUMANS/AND/SEX/OR/DISEASE/
/00193/
/METHOD/DETERM/BARIUM/BONE/HUMANS/RADIOACTIVATION ANALYSIS/
/00193/
/DETERM/STRONTIUM/BONE/HUMANS/RADIOACTIVATION ANALYSIS/7 UG PER G ASHED TISSUE/
/00193/
/DETERM/BARIUM/BONE/HUMANS/RADIOACTIVATION ANALYSIS/100 UG PER G ASHED TISSUE/
/00193/
/INCR/STRONTIUM/HUMANS/ADULTS/AGE 19-74/COMP W/CHILDREN/0-13/

FIGURA 46, PARTE 1

Comparação de minirresumo, resumo de autor e resumos publicados em *Chemical Abstracts* e *Biological Abstracts* (ver a parte 2 da figura)

Reproduzido de Lunin (1967) com permissão da Drexel University. O resumo do *Biochemical Journal* é reproduzido com permissão da Biochemical Society, Portland Scientific Press; o resumo do *Biological Abstracts*, com a permissão de BIOSIS; e o resumo do *Chemical Abstracts*, com permissão do Chemical Abstracts Service. Note-se que um resumo segue muito de perto o resumo de autor e o outro é simplesmente uma abreviação dele.

Craven (1987) analisa um método modular bastante diferente. Neste caso, um analista marca e codifica um texto para formar uma 'representação intermediária' que pode então ser usada, de modo semi-automático, para produzir resumos talhados às necessidades de diferentes públicos.

BA 32: 18857, 1958
18857. SOWDEN, ELEANOR M., and B. R. STITCH. (Med. Res. Council Radiobiol. Res.Unit, Atomic Energy Res.Establishment, Harwell, Didcot, Berks, Eng.) Trace elements in human tissue. 2. Estimation of the concentrations of stable strontium and barium in human bone. Biochem. Jour. 67(1): 104-109. 1957. -- A method is described for the determination of strontium and barium in human bone by radioactivation analysis. Results of analyses of 35 bone samples, from normal persons of both sexes and different ages, are given. The concentrations of Ba and Sr were of the order of 7 and 100 μg/g of ashed tissue respectively. No relationship between sex or disease of individuals with Sr and Ba concentration was noted. The concentration of Sr in the age group 1-13 years was significantly lower than that in the group 19-74 years. No significant difference was found in the concentrations of Sr and Ba in the various bones of those individuals examined. Results obtained in this survey are discussed and compared with those of other workers.-Auth. summ.

CA 51: 18184, 1957
II. Estimation of the concentrations of stable strontium and barium in human bone. Eleanor M. Sowden and S. R. Stitch. *Ibid.* 104-9.–A method based on the technique of Harrison and Raymond (*C.A.* 49, *12571g*) has been used for the detn. of Sr and Ba in human bone by radioactivation analysis. Results of analyses of 35 bone samples, from normal persons of both sexes and different ages, are given. The concns. of Ba and Sr were found to be of the order of 7 and 100 μg/g of ashed tissue, resp. No relation between sex or disease of individuals age group 0-13 yrs. was significantly lower than in the group 19-74 yrs. No significant difference was found in the concns. of Sr and Ba in the various bones of those individuals examined. The results obtained in this survey are discussed and compared with those of other workers.

Roland F. Beers, Jr.

FIGURA 46, PARTE 2

Minirresumos

O termo 'minirresumo' é bastante impreciso. Significaria simplesmente um resumo curto. Da forma como foi empregado por Lunin (1967), no entanto, o termo refere-se a um resumo altamente estruturado destinado essencialmente a buscas feitas em computador. Trata-se, com efeito, de um tipo de cruzamento entre um resumo e uma entrada de índice, e Lunin o define como um "índice-resumo legível por computador". Os termos utilizados no resumo são extraídos de um vocabulário controlado e reunidos numa seqüência especificada. Por exemplo, o enunciado "Existe um decréscimo da quantidade de zinco no sangue de seres humanos com cirrose do fígado" seria escrito assim:

/DECR/ZINCO/SANGUE/HUMANOS/CIRROSE/FÍGADO

Observe-se que o resumidor procura ater-se a uma seqüência de termos tão próxima quanto possível da estrutura normal da frase. O conteúdo de um docu-

mento pode ser descrito com algum detalhe por meio do emprego de uma série desses enunciados esquemáticos. Embora tenham sido imaginados basicamente para facilitar as buscas por computador, os minirresumos de Lunin também podem fazer sentido para o leitor inteligente. A figura 46, reproduzida do trabalho de Lunin, compara os resultados da técnica de minirresumos com resumos do *Biological Abstracts* e do *Chemical Abstracts* e com o resumo de autor.

Resumos telegráficos

A denominação 'resumo telegráfico' é também imprecisa. Ela implica uma representação de documento que é apresentada de modo muito lacônico: não com frases completas e semelhante a um telegrama. Na realidade, seria apenas uma cadeia de termos desprovida de sintaxe. Os minirresumos de Lunin são de estilo telegráfico. A expressão 'resumo telegráfico' foi empregada para designar um componente essencial do primitivo sistema de recuperação computadorizado desenvolvido na Western Reserve University (ver capítulo 11).

Capítulo 8

A redação do resumo

Assim como acontece com a indexação, só se aprende a ser um bom resumidor com a prática. O máximo que se pode fazer num livro como este é oferecer algumas diretrizes gerais.

E também como acontece na indexação, o bom resumidor aprenderá a ler/passar os olhos num documento para identificar rapidamente os pontos importantes. Cremmins (1996) trata, com detalhes, de como ler um artigo para captar os pontos mais importantes do modo mais eficiente possível e apresenta algumas regras com esta finalidade. Em grande parte isso é evidente por si mesmo e, de qualquer modo, indivíduos diferentes preferem técnicas diferentes para penetrar no âmago de um texto.

Em suma, as características de um bom resumo são brevidade, exatidão e clareza. O resumidor deve evitar redundância. O resumo deve, principalmente, ser estruturado a partir das informações contidas no título do item e não repetilas. Por exemplo, o título do artigo usado como exemplo nas figuras 3, 39 e 40 é "Pesquisa nacional de opinião pública sobre as atitudes norte-americanas acerca do Oriente Médio". A primeira linha de um resumo publicado desse artigo diz:

> Os resultados de uma pesquisa realizada em fevereiro de 1985 sobre as atitudes públicas norte-americanas acerca do Oriente Médio.

É claro que isso pouco acrescenta ao título, exceto a data. Note-se como os resumos das ilustrações 3, 39 e 40 partem do título sem repeti-lo.

O resumidor também deve omitir informações que o leitor provavelmente já conheça ou não lhe interessem diretamente. Isso inclui informações sobre antecedentes ou fatos de teor histórico, como, por exemplo, o motivo que levou à realização do estudo ou dados sobre a experiência da empresa que o executa. Borko e Bernier (1975) salientam que cabe ao resumidor indicar o que o autor fez e não o que tentou fazer, mas não conseguiu ou o que pretende fazer no futuro.

Quanto menor, melhor será o resumo, desde que o sentido permaneça claro e não se sacrifique a exatidão. Palavras desnecessárias como 'o autor' ou 'o artigo' são omitidas. Por exemplo, corta-se 'Este artigo examina...' para 'Examina...'. Abreviaturas e siglas convencionais são usadas sempre que for provável que os leitores as conheçam (por exemplo, OLP). Em outros casos, pode-se usar uma abreviatura desde que seu significado seja explicitado. Por exemplo:

> [...] no quadro da Cooperação Política Européia (CPE). As realizações [...] por parte da CPE [...]

Os resumos em alguns campos científicos chegam a empregar muitas abrevia-

turas. Apesar de economizar espaço, isso diminui a inteligibilidade e, realmente, exige mais tempo do leitor. A despeito da necessidade de brevidade, os resumos devem ser auto-suficientes; não se logrará um dos principais objetivos do resumo se o leitor tiver de consultar o original para entender o resumo!

É melhor evitar o jargão. As palavras de um jargão podem significar coisas diferentes para grupos diferentes de leitores e não ser compreendidas de maneira alguma por certas pessoas.

Alguns resumidores acham que devem mudar as palavras usadas pelo autor. Ainda que a paráfrase seja freqüentemente necessária para se obter brevidade, nada se tem a ganhar, na busca de originalidade, com a mudança das palavras empregadas pelo autor. Na realidade, é fácil distorcer o significado do original ao procurar, deliberadamente, por motivos estilísticos, encontrar expressões sucedâneas. Este aspecto é vigorosamente enfatizado por Collison (1971):

> É importante que o resumidor empregue, tanto quanto possível, o vocabulário do autor; a paráfrase é perigosa e pode conduzir o leitor a linhas de raciocínio que não eram aquelas pretendidas pelo autor (p. 11).

No entanto, Craven (1990) constatou que os resumos pouco empregam "seqüências literais de palavras dos textos completos", embora seu estudo fosse circunscrito a uma área temática muito restrita. O resumo é algo utilitário e não precisa ser uma obra de arte, embora Cremmins (1982) acredite que os resumos devam ter 'elegância' além de clareza e precisão.

A norma norte-americana sobre resumos (*Guidelines for abstracts*, 1997) especifica que os verbos devem ser usados na voz ativa (por exemplo, 'Os indicadores de função diminuem a revocação' e não 'A revocação é diminuída pelos indicadores de função') sempre que possível, mas que a passiva pode ser utilizada para 'enunciados indicativos e mesmo para enunciados informativos em que se deva destacar o receptor da ação'.* Esta restrição é muito imprecisa e é melhor esquecê-la: na maioria dos casos o tempo verbal preferido será óbvio por razões de estilo. Borko e Chatman (1963) e Weil (1970) sugerem que se empreguem os verbos no pretérito para a descrição de processos e condições experimentais e no presente para conclusões resultantes das experiências. O que é lógico: as atividades relatadas por um autor são coisas do passado, enquanto os resultados e as conclusões ainda pertencem ao presente. Borko e Bernier (1975) são mais explícitos ao recomendar a voz ativa e o pretérito para resumos informativos, e a voz passiva e o presente para resumos indicativos.

Até hoje foram elaborados muitos conjuntos de regras sobre redação de resumos. Talvez o conjunto mais conciso de princípios destinados à elaboração de resumos seja o produzido pelo Defense Documentation Center (1968), reproduzido na figura 47. Em poucos e breves enunciados sintetiza as regras adotadas pelo centro sobre o que incluir, o que não incluir, qual a extensão que o resumo deve ter e qual o tipo de terminologia a ser adotado. Uma exposição mais completa, mas também concisa, encontra-se em relatório de Payne et al. (1962), e é reproduzida no apêndice 1 deste livro.

* A norma brasileira sobre resumos — NBR 6028, da Associação Brasileira de Normas Técnicas (ABNT) — também preceitua o emprego da voz ativa, sem fazer menção ao uso da voz passiva (N.T.)

ESQUEMA

Sucintamente:

1. Sempre um resumo informativo, se possível
2. 200–250 palavras
3. A mesma terminologia técnica do documento
4. Conteúdo
   a. Objetivos ou finalidade da pesquisa
   b. Métodos da pesquisa
   c. Resultados da pesquisa
   d. Validade dos resultados
   e. Conclusões
   f. Aplicações
5. Algarismos para números, quando possível
6. Frases em lugar de orações, palavras em lugar de frases, quando possível
7. Nenhum símbolo ou caráter não-convencional ou raro
8. Nenhuma abreviatura incomum
9. Nenhuma equação, nota de rodapé, preliminares
10. Nenhum dado de catalogação descritiva
11. Classificação de sigilo
12. Controles de disseminação, se houver
13. Revise-o.

FIGURA 47
Princípios para redação de resumos, do Defense Documentation Center (1968)

Conteúdo e formato

O que se deve incluir num resumo depende muito, é claro, do tipo de publicação que se tem em mira. Um longo resumo indicativo de um tipo de relatório de pesquisa mencionaria os objetivos da pesquisa, os procedimentos experimentais e de outra natureza adotados, os tipos de resultados obtidos (um resumo informativo conteria os próprios resultados, pelo menos de forma condensada), e as conclusões do autor quanto à importância dos resultados. O tratamento a ser dado a um artigo de história, por outro lado, seria bem diferente. O resumo, por exemplo, daria ênfase à tese ou conclusões do autor, tomando o cuidado de mencionar os períodos, localidades geográficas e personalidades envolvidos.*

Em áreas temáticas especializadas, o resumidor pode receber instruções

* Tibbo (1992) mostrou que as normas publicadas relativas à redação de resumos são muito mais pertinentes às ciências do que às humanidades.

sobre certas coisas a serem procuradas nos artigos e destacá-las com clareza nos resumos. Isso pode incluir itens tão diversos quanto dosagem de um medicamento, condições climáticas, idade dos indivíduos, tipos de solo, equações empregadas ou o elemento componente de uma liga. Os resumos costumam ser de redação mais fácil quando o conteúdo temático trata de objetos concretos, e são de redação mais difícil quanto mais abstrato ou nebuloso for o assunto.

A maioria dos resumos é apresentada no formato convencional de referências bibliográficas seguidas do texto do resumo. Em algumas publicações, no entanto, o resumo precede a referência bibliográfica, e sua primeira linha é realçada de alguma forma, como no exemplo seguinte:

> A MIGRAÇÃO DE MÃO-DE-OBRA DE MOÇAMBIQUE PARA AS MINAS DA ÁFRICA DO SUL continua sendo um elemento importante nas relações econômicas entre estes países....
>
> Brockmann, G. Migrant labour and foreign policy: the case of Mozambique. *Journal of Peace Research*, 22, 1985, 335-344.

Esta é uma forma de apresentação mais atraente, muito parecida com o cabeçalho de uma matéria de jornal, e que pode captar a atenção do leitor sem grande esforço. Weil et al. (1963) referem-se a isso como resumo 'orientado para o leitor', resumo 'de tópico frasal em primeiro lugar' ou resumo 'orientado para resultado' (embora o título não tenha de ser necessariamente relacionado aos resultados). Se for adequado, o título do artigo poderá transformar-se nesse cabeçalho, vindo em seguida um tópico frasal que o desenvolva.

Considera-se um resumo completo como sendo composto de três partes: a *referência*, que identifica o item resumido; o *corpo* do resumo (o texto); e a *assinatura*. Este último elemento é a atribuição da origem do resumo: as iniciais do resumidor ou a indicação de que o resumo foi elaborado pelo autor do item, de que se trata de um resumo modificado de autor, ou deriva de uma fonte diversa, como, por exemplo, outro serviço de resumos.

Muitos resumos parecem situar-se na faixa de 100–250 palavras, mas, como se disse antes, é natural que a extensão varie de acordo com certos fatores, como o tamanho do próprio documento, o alcance de seu conteúdo temático, a importância que lhe é atribuída, sua disponibilidade física e acessibilidade intelectual (por exemplo, itens de difícil localização, como trabalhos apresentados em eventos, ou em línguas pouco conhecidas, seriam resumidos com mais detalhes do que outros itens). Borko e Bernier (1975) sugerem que os resumos da literatura científica deveriam ter comumente entre um décimo e um vigésimo da extensão do original, embora Resnikoff e Dolby (1972) indiquem que um trigésimo talvez seja mais comum.

Borko e Bernier (1975) nos dão um conselho útil para a seqüência do conteúdo:

> O corpo do resumo pode ser ordenado de modo a poupar o tempo do leitor. A colocação das conclusões em primeiro lugar satisfaz ao leitor e poderá dispensá-lo de continuar a leitura. Ele pode aceitar ou rejeitar as conclusões sem que precise conhecer os resultados em que se basearam. O desenvolvimento das informações virá em último lugar. Verificou-se ser desnecessário rotular cada parte do resumo, como, por exemplo, *conclusões, resultados* ou *métodos*; normalmente os leitores sabem qual é a parte que estão lendo. A ordenação das partes do corpo do resumo é feita com a mesma finalidade com que se organizam as partes de uma matéria de jornal — para comunicar a informação de modo mais rápido.
>
> Não convém abrir parágrafos. O resumo é breve; deve exprimir um raciocínio homogêneo e ser redigido como um único parágrafo (p. 69).

De fato, a tendência recente tem sido no sentido de dividir os resumos em pedaços menores mediante a abertura de parágrafos e até mesmo o uso de entretítulos. Isso tem sido verificado principalmente em periódicos de medicina. Um exemplo, da própria literatura de ciência da informação, e que estuda esta mesma situação, é mostrado na figura 48. Esse tipo de resumo passou a ser conhecido como 'resumo estruturado', embora a forma como esta expressão seja aí empregada seja bastante diferente da forma como a utilizo.

Curiosamente, desde 1988, é provável encontrar na literatura médica um número maior de artigos sobre 'resumos' do que na literatura de ciência da informação. Resumos 'estruturados' de artigos médicos foram publicados pela primeira vez na revista *Annals of Internal Medicine*, que solicitava aos autores que preparassem os resumos conforme um formato que lhes era prescrito, tendo sido definidas regras bastante apuradas para sua redação (ver, por exemplo, Haynes et al., 1990). A figura 49 contém uma síntese do tipo de informação a ser incluída, mas as instruções aos autores são muito mais detalhadas.

Não foi sem polêmica que os resumos estruturados foram introduzidos nos periódicos de medicina. Haynes et al. (1990) sugerem que a formatação muito rígida pode estimular alguns autores a reivindicar mais do que seria cabível. Por exemplo, se houver um entretítulo *método* ou *delineamento experimental*, para prender a atenção do leitor, isso pode levar alguns a alardear um enfoque mais rigoroso do que o que seria realmente justificável.

Froom e Froom (1993a,b) mostraram que os resumos estruturados dos *Annals of Internal Medicine* nem sempre continham todas as informações exigidas nas instruções para os autores, mesmo quando as informações solicitadas estavam presentes no próprio artigo. Haynes (1993) critica esse estudo, mas sua crítica não é convincente. Taddio et al. (1994), baseando-se em estudo mais amplo, cobrindo 300 resumos extraídos de três periódicos, verificou que os resumos estruturados apresentavam maior probabilidade de conter informações mais completas de importância para a pesquisa do que os resumos não-estruturados. Os aspectos sobre avaliação serão tratados no próximo capítulo.

Mesmo que os resumos estruturados desse tipo possam ter seus méritos, muitas vezes suas pretensões são exageradas. Por exemplo, Haynes et al. (1990) alegam que eles "podem facilitar a avaliação pelos pares antes da publicação, ajudar os leitores que exercem a clínica a encontrar artigos que sejam tanto cientificamente corretos quanto aplicáveis à prática profissional, além de

permitir buscas bibliográficas informatizadas mais precisas", embora nem todas essas alegações sejam documentadas.

RESUMO

ANTECEDENTES: Os resumos estruturados, que, como este, contêm vários entretítulos, substituíram os resumos tradicionais na maioria dos periódicos médicos. Estudos de avaliação mostraram que esses resumos normalmente oferecem mais informações, são de melhor qualidade, facilitam a avaliação pelos pares e, em geral, são bem-aceitos.
OBJETIVO: O objetivo dos estudos aqui reportados foi investigar uma outra possível vantagem dos resumos estruturados, a saber, se neles as buscas são ou não são mais fáceis de executar.
MÉTODO: São relatados dois estudos. No estudo 1, efetuado numa base de dados eletrônica, solicitou-se a 52 leitores que encontrassem as respostas a duas perguntas feitas a cada um de oito resumos em um formato (digamos, tradicional) seguidas de duas questões para cada um de oito resumos compostos no outro formato. Foram automaticamente registrados os dados de tempo e erros. No estudo 2, efetuado numa base de dados impressa, solicitou-se a 56 leitores que encontrassem cinco resumos que relatassem determinado tipo de estudo (por exemplo, estudos com escolares e testes de leitura) e depois achassem mais cinco outros que relatassem outro tipo de estudo. Além disso, a ordem e apresentação do formato foram compensadas. Os dados de tempo e erro foram registrados manualmente.
RESULTADOS: No estudo 1, os participantes tiveram desempenho significativamente mais rápido e cometeram significativamente menos erros com os resumos estruturados. Houve, contudo, alguns inexplicáveis efeitos da prática. No estudo 2, os participantes novamente tiveram desempenho significativamente mais rápido e cometeram significativamente menos erros com os resumos estruturados. No estudo 2, contudo, houve efeitos de transferência assimétricos: participantes que responderam primeiro aos resumos estruturados responderam mais rapidamente aos resumos tradicionais seguintes do que o fizeram os participantes que responderam primeiro aos resumos tradicionais.
CONCLUSÕES: Em geral, os resultados, apesar de certas ressalvas, apóiam a hipótese de que é mais fácil para os leitores fazer buscas em resumos estruturados do que em resumos tradicionais.

FIGURA 48

Exemplo de resumo altamente formatado

Reproduzido de Hartley et al. (1996) com permissão do *Journal of Information Science*

É interessante que, quase na mesma época em que a literatura médica descobria esse tipo de resumo, Trawinski (1989) examinava métodos similares de redação de resumos em ciência da informação. Ele também comparou as características dos resumos assim redigidos com resumos da base de dados INSPEC.

A literatura sobre resumos estruturados continua a crescer. Hartley (1998) defende a mais ampla adoção desses resumos em periódicos científicos. Ele (Hartley, 2000b) também argumenta que é preciso alguma forma de resumo estruturado junto às revisões sistemáticas da literatura médica. Alega que tais resumos devem ser mais fáceis de ler do que os resumos de artigos de pesquisa médica porque as revisões sistemáticas têm como alvo um público mais amplo.

Artigos originais

1. Objetivo: a(s) questão(ões) exata(s) abordada(s) pelo artigo
2. Delineamento experimental: o delineamento básico do estudo
3. Ambiente: a localização e o nível da assistência clínica
4. Pacientes ou participantes: o modo de seleção e o número de pacientes ou participantes que iniciaram e chegaram ao fim do estudo
5. Intervenções: o tratamento ou intervenção exata, se houve algum
6. Principais medidas de resultado: a medida fundamental de resultado do estudo planejada antes de iniciada a coleta de dados
7. Resultados: os principais achados
8. Conclusões: as principais conclusões inclusive aplicações clínicas diretas.

Artigos de revisão

1. Finalidade: o objetivo fundamental do artigo de revisão
2. Fontes de dados: um apanhado sucinto das fontes dos dados
3. Seleção dos estudos: o número de estudos selecionados para a revisão e como foram selecionados
4. Extração dos dados: regras para o resumo dos dados e como foram aplicadas
5. Resultados da síntese de dados: os métodos de síntese de dados e principais resultados
6. Conclusões: conclusões principais, inclusive aplicações potenciais e necessidade de pesquisas adicionais

FIGURA 49

Informações essenciais de que necessitam os clínicos para avaliar a relevância e a qualidade de artigos e, portanto, para sua inclusão em resumos estruturados

Reproduzido de Haynes et al. (1990) com permissão dos *Annals of Internal Medicine*

Uma das objeções aos resumos estruturados, expressa por editores de periódicos, é que ocupam mais espaço. Essa questão foi estudada por Hartley (2002). Ele concluiu que os resumos estruturados realmente ocupam mais espaço (normalmente seu tamanho é 21% (às vezes mais) maior do que os resumos tradicionais), porém isso somente afetaria aquelas revistas (relativamente raras) em que os artigos se sucedem um em seguida ao outro e não os periódicos em que cada artigo abre uma nova página.

Os tipos de erros mais evidentes que ocorrem na indexação de assuntos também ocorrem na redação de resumos: aspectos que deveriam ser incluídos não o são, e outros que são incluídos ficariam melhor se fossem omitidos. Também podem ocorrer erros de transcrição, principalmente quando se trata de fórmulas ou valores numéricos. Sempre conferir e submeter à revisão editorial por parte de alguém mais experiente o trabalho de resumidores inexperientes. Borko e Bernier (1975) confirmam a utilidade de um bom editor de resumos:

> Os editores de resumos parecem desenvolver um sexto sentido que os faz saber quando está faltando uma parte importante do conteúdo. Eles procuram, e esperam encontrar, certas categorias de informação, como os métodos e equipamentos utilizados, os dados coletados e as conclusões (p. 12).

Um serviço de resumos provavelmente adotará algumas diretrizes sobre certos pontos, tais como ortografia, pontuação e uso de maiúsculas. Como isso cons-

titui, em grande parte, uma questão de preferência individual, parece despropositado apresentar exemplos.

Para ajudar o resumidor em seu trabalho, principalmente num programa de treinamento, convém preparar algum tipo de planilha que o oriente sobre aquilo que deve procurar numa publicação. Uma planilha* como essa incluiria certos aspectos, como, por exemplo:

> *Tipo e objetivo* [Tipo de estudo, se experimental, teórico, de revisão, pesquisa básica ou aplicada, desenvolvimento. Objetivo: uma proposição do problema, uma definição do que exatamente é pesquisado.]
> *Plano experimental ou modelo teórico* [Características importantes, novos enfoques, hipótese a ser comprovada, resultados esperados quando o trabalho foi iniciado. O que torna este trabalho diferente, tanto experimental quanto analiticamente, do trabalho de outros pesquisadores?]
> *Condições estudadas* [Parâmetros variados, limites envolvidos, controles impostos.]
> *Procedimentos* [Técnicas novas empregadas, transformações utilizadas ou desenvolvidas, como os resultados foram obtidos.]
> *Pressupostos* [Quais os pressupostos diretos e indiretos, e são eles convencionais?]
> *Conclusões principais* [Principais conclusões do autor, outras conclusões apoiadas nos dados, resultados negativos importantes.]
> *Conclusões secundárias* [Pontos de menor importância ou aqueles de áreas periféricas da pesquisa podem ser relatados se forem julgados suficientemente úteis. Podem ser apresentadas interpretações e inferências e extrapolações razoáveis. Não são convenientes associações teóricas imprecisas e questões conjeturais.]
> *Importância ou utilidade* [Importância e competência do trabalho realizado. Aplicações potenciais.]
> *Limitações e deficiências* [As hipóteses são indevidamente restritivas ou limitantes? O modelo teórico está muito distante de possível aplicação prática? Há falhas técnicas? O enfoque do problema impôs limitações aos resultados? Que grau de complexidade foi adotado? Houve análise suficiente dos dados, principalmente quanto a possíveis erros?]
> *Comentários críticos* [Eventual erro fundamental e magnitude dos erros. Eventual publicação anterior desta informação. Existem pesquisas similares e qual é a posição que o presente trabalho ocupa na bibliografia? Quais as características que são particularmente meritórias? A interpretação dos resultados é razoável?]

É claro que nem todas essas categorias serão aplicáveis a todo item a ser resumido e as três últimas somente a resumos críticos. Solov'ev (1971) estuda o uso, na redação de resumos, deste método baseado em questionário.

Hoje, naturalmente, é provável que alguma forma de auxílio ao processo de redação do resumo, em linha e interativo, seja mais atraente do que a adoção desse tipo de enfoque estruturado, ainda que exibido em linha na tela. Craven (1996) desenvolveu um protótipo de sistema destinado a assistir os resumidores e chegou a testá-lo pelo menos em caráter preliminar. O auxílio à redação de resumos inclui um tesauro como um dos componentes (Craven, 1993).

Alguns autores procuraram desenvolver diretrizes para redação de resumos de certos tipos de documentos. Por exemplo, Solov'ev (1981) sugere que resumos de teses de doutorado focalizem os seguintes pontos: importância atual do assunto, problema tratado e objetivo da pesquisa, novidade científica, metodologia, resultados e conclusões (inclusive implementação dos resultados).

* Os títulos e descrições da planilha aqui exemplificada baseiam-se nos utilizados no projeto de resumos modulares de Herner and Company (Lancaster et al., 1965).

Embora de modo um tanto confuso e, por isso, com trechos de difícil interpretação, o Centro de Documentação sobre Refugiados do Alto Comissariado das Nações Unidas para os Refugiados (UNHCR) condensou os fundamentos da redação de resumos num único diagrama (figura 50). São particularmente úteis os critérios de avaliação à esquerda do diagrama. Note-se que o resumo deve ser avaliado com base em sua linguagem e conteúdo, sua obediência ao 'estilo da casa' (extensão, estrutura, convenções ortográficas e de pontuação) e, o que é mais importante, o grau com que ele atende às necessidades do usuário.

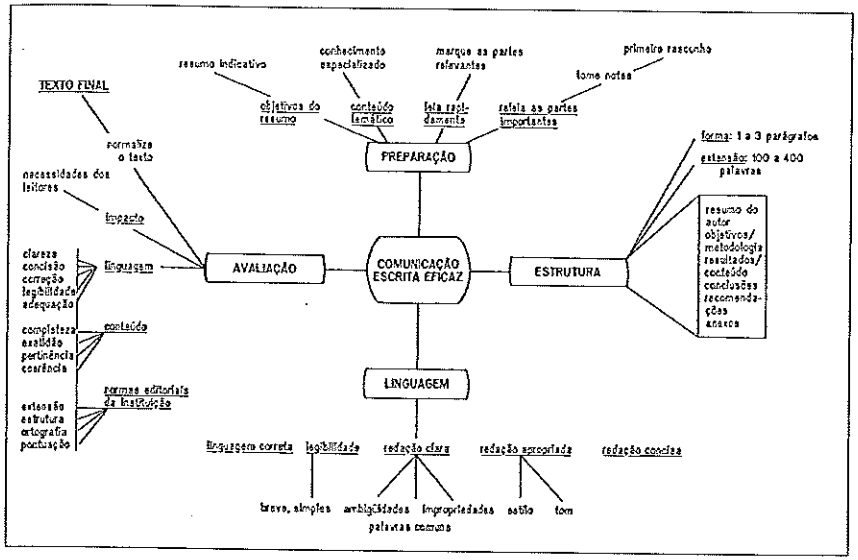


FIGURA 50

Fundamentos da redação de resumos

Reproduzido do UNHCR Refugee Documentation Centre (1985), com permissão do Centro de Documentação sobre Refugiados do Alto Comissariado das Nações Unidas para os Refugiados

Foram desenvolvidos 'modelos' mais formais do processo de redação de resumos (por exemplo, por Karasev, 1978). Embora tais modelos contribuam para nosso entendimento das etapas percorridas intuitivamente pelo resumidor, parecem ser de pouco valor prático para os resumidores.

Mesmo que se reconheçam alguns princípios gerais implícitos no trabalho de resumir, é claro que cada resumidor terá seu próprio modo de implementá-los. Com efeito, Endres-Niggemeyer (1994, 1998) admitiu ter identificado 453 estratégias diferentes, com base na gravação em fita dos protocolos que apenas seis resumidores pensavam em voz alta.

Os aspectos teóricos dos resumos e sua redação são abordados mais amiúde

por autores europeus, principalmente Endres-Niggemeyer (1998) e Pinto. Esta autora apresenta, num livro em espanhol (Pinto, 2001), um estudo completo sobre redação de resumos em seu contexto lingüístico mais amplo. Encontram-se versões parciais em Pinto (1995) e Pinto e Lancaster (1999). Ela também analisou o papel da semiótica, lógica e psicologia cognitiva na análise de conteúdos textuais (Pinto, 1994). Pinto e Gálvez (1999) analisam a redação de resumos em termos de paradigmas comunicacionais, físicos, cognitivos e sistêmicos.

Resumidores

Os resumos podem ser preparados pelos próprios autores dos documentos, por outros especialistas dos respectivos assuntos ou resumidores profissionais. Muitos periódicos científicos exigem que os autores preparem resumos que acompanhem seus artigos. É crescente o aproveitamento destes resumos pelos serviços de resumos impressos, que assim não precisam redigi-los de novo.

Igual ao que acontece na indexação, o resumidor precisa ter conhecimento do conteúdo temático tratado no documento, embora não precise ser um especialista. Requisito importante é a capacidade de redigir e editorar textos, e um trunfo precioso para o resumidor é a aptidão de ler e compreender com rapidez.

Borko e Bernier (1975) advertem que os autores não necessariamente redigem os melhores resumos dos próprios artigos. Os autores comumente não possuem formação e experiência em redação de resumos, bem como carecem do conhecimento das regras adotadas nessa atividade. As publicações de resumos mais prestigiosas comumente conseguem recrutar como resumidores especialistas numa área, que aceitam o encargo de redigir resumos em sua especialidade sem remuneração ou a troco de honorários modestos. Borko e Bernier afirmam que: "Quem aprendeu a redigir resumos e também é especialista num campo do conhecimento redige os melhores resumos", afirmação contra a qual é difícil argumentar. Como os especialistas comumente são voluntários, talvez seja difícil deles obter pontualidade na redação dos resumos. O resumidor profissional custa caro, mas é pontual e pode fazer um excelente trabalho quando lida com áreas do conhecimento que não lhe sejam totalmente desconhecidas.

Atualmente autores e editoras contam com poucos incentivos para 'enfeitar' os resumos de modo a tornar o trabalho que isso envolve mais atraente do que de fato é. Price (1983) argumentou que isso poderia tornar-se um perigo num ambiente completamente eletrônico (ver capítulo 16). As editoras gostariam de estimular o uso dos resumos porque provavelmente seriam remuneradas com base nisso. Os autores gostariam de ampliar sua utilização, se este fator se tornasse, o que não seria impossível, um critério a ser adotado em decisões relativas à promoção e estabilidade no emprego. O vocábulo *spoofing* ou *spamming* foi usado para designar o enfeitamento de páginas da Rede de modo a aumentar sua recuperabilidade (capítulo 16).

A redação de resumos e a indexação são atividades intimamente relacionadas e há fortes razões para que sejam reunidas. É pequena a distância entre a etapa de análise conceitual da indexação e a preparação de um resumo aceitável. Além disso, a disciplina adicional imposta pela redação do resumo ajuda na decisão sobre o que deve ser incluído e o que pode ser omitido na indexação. O fato de ambas as atividades envolverem uma combinação de leitura e passar de olhos é outro motivo pelo qual é eficiente reuni-las, delas se incumbindo uma mesma pessoa, sempre que isto for viável.

Qualidade e coerência na redação de resumos

Jamais dois resumos do mesmo documento serão idênticos se forem redigidos por pessoas diferentes ou pela mesma pessoa em momentos diferentes: o conteúdo descrito será o mesmo, mas a forma de descrevê-lo será diferente. A qualidade e a coerência são um pouco mais vagas quando se aplicam a resumos do que quando se aplicam à indexação. Aparentemente existem duas facetas principais concernentes à qualidade:

1. Os 'argumentos' essenciais do documento são postos em relevo no resumo?
2. Esses argumentos são descritos exata, sucinta e inequivocamente?

Em certa medida, portanto, a qualidade da redação de resumos é aferida segundo critérios que são muito similares aos adotados na avaliação da indexação. A primeira etapa da redação de resumos corresponde, realmente, como na indexação, à análise conceitual — quais os aspectos a realçar? — e a segunda etapa corresponde à tradução dessa análise conceitual em frases (normalmente).

A qualidade da análise conceitual pode ser aferida, provavelmente, em cotejo com as instruções relativas a conteúdo baixadas pela instituição para a qual o resumo é redigido. Por exemplo:

1. Foram incluídos o âmbito e a finalidade do trabalho?
2. Os resultados foram indicados ou resumidos?
3. As conclusões do autor foram resumidas?

   e assim por diante.

Poder-se-á, então, avaliar a coerência entre dois resumos, neste nível conceitual, no que tange ao grau com que os resumidores estiveram de acordo quanto aos pontos a incluir.

A avaliação da qualidade da fase de 'tradução', no processo de elaboração do resumo, é um pouco mais complicada, porque exatidão, ambigüidade e brevidade são critérios um tanto subjetivos. Um resumidor experiente poderá aplicá-los, porém, ao julgar o trabalho de pessoas menos experientes. Não deve haver grande preocupação com a coerência na fase de tradução do processo de elaboração do resumo: existe a possibilidade de tratar a mesma questão de várias formas diferentes, cada uma delas exata e inequívoca e, talvez, igualmente sucinta.

	Texto integral	*Resumos*
Número de itens julgados relevantes	12	15
Número de itens julgados não-relevantes	38	35
Total	50	50

FIGURA 51
Resultados hipotéticos de um teste de previsibilidade de relevância

O teste definitivo de um bom resumo consiste simplesmente em perguntar: 'será que ele permite ao leitor prever com exatidão se um item resumido é ou não relevante para seus interesses atuais?' No que tange a um leitor específico e uma necessidade de informação específica, é possível testar isso com base, digamos, em 50 resumos copiados na impressora em resposta a uma busca em linha. Os resultados do estudo seriam os mostrados na figura 51: os resumos sugeriam que 15 itens seriam relevantes, mas apenas 12 se confirmaram como tal. Além disso, se se constatar que nem todos os 12 considerados relevantes a partir do texto foram também julgados relevantes a partir dos resumos, estes teriam falhado de ambos os modos: sugeriram que alguns itens eram relevantes quando não o eram, e que outros não eram relevantes quando o eram.

Esse tipo de estudo, naturalmente, é um tanto difícil de fazer. Ademais, seus resultados aplicam-se apenas a determinado usuário e determinada necessidade de informação; mude-se o usuário ou a necessidade de informação e os resultados mudarão. A maioria dos usuários de serviços de resumos, ou bases de dados em linha, já terá passado pela situação, talvez com relativa freqüência, quando um resumo desperta o apetite por um documento que acaba sendo muito diferente do que se esperava. Então, os resumos frustraram esses usuários, embora talvez tivessem atendido de modo bastante adequado a outros usuários.

A utilidade dos resumos na previsão da relevância de documentos para determinado usuário será examinada com mais detalhes no capítulo seguinte. Embora tenham sido realizados vários estudos de previsibilidade de relevância, são muito poucas as pesquisas sobre as atividades dos resumidores. Com efeito, já foram realizados mais estudos sobre avaliação de extratos do que de resumos. Por exemplo, tanto Rath et al. (1961b) quanto Edmundson et al. (1961) descobriram que as pessoas não eram muito coerentes (com os outros ou consigo mesmas) ao selecionar de um texto as frases que consideravam como os melhores indicadores de seu conteúdo.

Edmundson et al. (1961) sugerem vários métodos de avaliação de resumos:

1. Julgamento intuitivo, subjetivo;
2. Comparação com um resumo 'ideal';
3. Determinação da medida em que perguntas de teste sobre um documento podem ser respondidas pelo resumo;
4. Recuperabilidade do documento pelo resumo.

É claro que os resumos são avaliados pelos editores e outras pessoas que trabalham nos centros de informação ou nas editoras, provavelmente utilizando o método intuitivo. É provável que, quanto mais se utilizar a busca em texto livre em lugar da indexação feita por seres humanos, crescerá a importância do método de avaliação baseado na 'recuperabilidade'. Os critérios para aferir a 'recuperabilidade' de um resumo não são necessariamente os mesmos utilizados para avaliá-lo com base na previsibilidade da relevância (ver as considerações sobre 'Questões de compatibilidade').

Vinsonhaler (1966) propõe métodos comportamentais para avaliar a qualidade de resumos com base na 'validade de conteúdo' ou 'validade previsível'. Num estudo de validade de conteúdo, os sujeitos julgam o grau com que o documento e o resumo são 'similares', empregando talvez uma escala de similaridade de sete pontos. Alternativamente, pode-se aplicar um teste para determinar em que medida um resumo discrimina documentos, especialmente quando seu conteúdo temático é bastante similar. Vinsonhaler propõe, para medir a discriminabilidade, um teste em que os sujeitos examinam um documento e em seguida procuram identificar o resumo correspondente num folheto que contém resumos. Um teste de validade previsível determina em que medida decisões tomadas sobre similaridade dos resumos coincidem com decisões quanto à similaridade tomadas com base nos próprios documentos: se os resumos forem 'bons', grupos de resumos com base na similaridade devem coincidir com grupos de documentos com base na similaridade. O segundo teste de validade previsível é mais convencional: determina-se a medida em que os resumos prevêem corretamente a relevância dos documentos. Vinsonhaler sugere um teste de cruzamento em que um grupo de pessoas avalia a relevância de um conjunto de documentos para um enunciado de pedido de busca e, em seguida, depois de um intervalo de tempo adequado, faz o mesmo com os resumos dos documentos. O segundo grupo de pessoas procede em seqüência inversa, primeiro avaliando os resumos e depois os documentos.

Mathis (1972) propôs que os resumos fossem avaliados com base num 'coeficiente de dados' (*CD*). O *CD* é expresso pela fórmula C/L, onde C é um 'fator de conservação de dados' e L um 'fator de conservação de extensão'. C é uma medida do grau com que todos os 'conceitos' (Mathis refere-se a eles como 'elementos de dados') do documento são conservados no resumo. L é simplesmente o número de palavras do resumo dividido pelo número de palavras do documento. O *CD* é um valor numérico, e, quanto maior o valor, melhor. Ele favorece a concentração e a compressão: capacidade de conservar todos os elementos essenciais do texto com o mínimo de palavras. Melhora-se o valor aumentando-se a quantidade de elementos de dados presentes ou reduzindo-se a quantidade de palavras no resumo. Mathis sugere que um valor de *CD* inferior à unidade indicaria um resumo de qualidade inaceitável. Trata-se de método criativo, embora dependa totalmente da capacidade de identificar 'elementos de dados'. Mathis propõe que sejam identificados mediante critérios sintáticos.

Assim como qualquer outro tipo de texto, os resumos podem ser avaliados com base na 'legibilidade', empregando-se fórmulas clássicas de legibilidade. Dronberger e Kowitz (1975) adotaram a fórmula de facilidade de leitura, de Flesch, para comparar resumos de *Research in Education* com os relatórios correspondentes, e verificaram que os níveis de legibilidade eram significativamente baixos, provavelmente porque careciam de redundância. Também King (1976), adotando um critério 'cloze',* observou que os resumos de *Child Development Abstracts* eram menos legíveis do que os itens em que se baseavam.

Hartley (1994) aplicou tanto os escores de legibilidade de Flesch** quanto testes cloze (compreensão) na comparação de quatro versões diferentes dos mesmos resumos. Sua conclusão foi que os resumos poderiam ser melhorados (isto é, redigidos de modo mais claro) se fosse mudado o tamanho do tipo, se fosse estruturado (em parágrafos com entretítulos) e se fossem reescritos. Posteriormente, Hartley e Sydes (1996) estudaram as preferências do leitor quanto à disposição gráfica dos resumos estruturados.

Hartley (2000) identifica três fatores que influem na clareza dos resumos: linguagem (legibilidade), a maneira como as informações são apresentadas (seqüencial ou estruturada) e o estilo gráfico. Também descreve diferentes soluções gráficas para a apresentação dos resumos estruturados.

Salager-Meyer (1991) analisou, de uma perspectiva lingüística, uma amostra de resumos de artigos médicos, tendo chegado à conclusão de que metade deles era 'mal-estruturada' (isto é, possuía deficiência de discurso). Uma vez que 'deficiência de discurso' pode incluir coisas do tipo dispersão conceitual (por exemplo, resultados relatados em diferentes lugares do resumo) como também omissão de um elemento importante (por exemplo, o objetivo da pesquisa) do resumo, o autor sugere que os resumos que padeçam desses defeitos serão menos eficientes na transmissão de informações.

Borko e Bernier (1975) apresentaram aquela que talvez seja a lista mais abrangente de possíveis critérios de avaliação de resumos, como se vê a seguir:

1. Uma classificação global de qualidade (atribuída por avaliadores humanos).
2. A medida em que a norma NISO (ANSI) ou outra norma é respeitada (que é também tida como um componente principal do método de avaliação recomendado por Mathis (1972).***

* Técnica e teste de leitura em que, num texto, são omitidas palavras propositalmente e segundo um padrão definido (por exemplo, toda quinta palavra). Os espaços em branco deixados serão preenchidos pelos sujeitos que estiverem sendo avaliados. (N.T.)

** O escore de facilidade de leitura de Flesch [Flesch Reading Ease (R.E.)] considera a extensão das frases e a extensão das palavras no texto. A fórmula original é R.E. = 206,835 - 0,846w - 1,015s (onde w é o número médio de sílabas em 100 palavras e s é o número médio de palavras por frase. Os escores normalmente situam-se na faixa 0–100 em que os valores menores refletem maior dificuldade (Hartley, 2000c).

*** Ver também, porém, os comentários sobre normas no capítulo 9.

3. A inclusão de informações importantes e a exclusão de informações sem importância.
4. Ausência de erros.
5. Coerência de estilo e legibilidade.
6. Previsibilidade da relevância.
7. Capacidade de servir como substituto do original (resumos informativos).
8. Adequação como fonte de termos de indexação.

Esta lista, evidentemente, representa vários níveis de critérios. Por exemplo, todos os critérios do terceiro ao quinto provavelmente seriam levados em conta em qualquer classificação 'global'. Um método de avaliação da medida em que um resumo pode servir em lugar do original (critério 7) consiste em comparar a capacidade de grupos de indivíduos responderem a questões baseadas em: a) nos resumos, e b) no texto integral. Payne et al. (1962) relataram estudos desse tipo.

Com efeito, os estudos de Payne englobavam três métodos de avaliação diferentes:

1. *Coerência.* Foram utilizados especialistas de assunto para comparar resumos com base na similaridade da quantidade de informações apresentadas.
2. *A quantidade de redução de texto obtida.*
3. *Utilidade.* Os estudantes responderam a questões técnicas baseadas nos artigos de sua área de especialização. Alguns deles liam os artigos, outros apenas os resumos. As respostas dos dois grupos eram comparadas. Este método foi também usado por Hartley et al. (1996) para comparar diferentes tipos de resumos: a conclusão foi que os resumos estruturados (formatados em diferentes parágrafos, cada um com seu entretítulo) podiam ser usados de modo mais eficiente.

No programa TIPSTER (ver capítulo 14), são empregados dois métodos de avaliação de resumos: 1) utilização do resumo para julgar a relevância dos documentos, e 2) utilização do resumo como base para a classificação dos documentos (isto é, classificação baseada no resumo em comparação com a classificação baseada nos textos completos).

As avaliações da qualidade dos resumos publicadas nos últimos anos têm se concentrado, na maior parte, em resumos estruturados. Hartley e Benjamin (1998) compararam resumos tradicionais e estruturados redigidos por autores de artigos submetidos a quatro periódicos britânicos de psicologia. Estudantes de psicologia participaram da avaliação. Os resumos estruturados foram julgados significativamente mais legíveis, significativamente mais longos e significativamente mais informativos.

Poucos trabalhos foram realizados para avaliar resumos publicados em comparação com os textos a que se referem. No entanto, um útil estudo desse tipo foi relatado por Pitkin et al. (1999). Eles avaliaram, dessa forma, 88 resumos publicados em seis importantes revistas médicas. Os resumos eram consi-

derados 'deficientes' quando incluíam dados diferentes dos dados constantes do próprio artigo ou deixavam de incluir dados por completo. Com base nesses critérios, uma quantidade significativa de resumos foi considerada deficiente, cerca de 18% na revista que correspondia ao melhor caso e 68% no pior caso.

Em decorrência desse estudo o *JAMA* (*Journal of the American Medical Association*) introduziu um programa de melhoria de qualidade (Winkler, 1999). Foram adotados os seguintes critérios:

1. Os entretítulos do resumo são coerentes com o formato de resumo estruturado.
2. Os dados no resumo são coerentes com o texto, tabelas e figuras.
3. Os dados ou informações do resumo estão presentes no texto, tabelas ou figuras.
4. Fornece os anos de estudo e a duração do acompanhamento.
5. Os resultados das Principais Medidas de Resultados são apresentados na seção de Resultados (evitar informações seletivas).
6. Os resultados são quantificados com numeradores, denominadores, *odds ratios* [razões de chances, razões de diferenças] e intervalos de confiança, onde isso for apropriado.
7. Apresentam-se, sempre que possível, diferenças absolutas e não diferenças relativas (por exemplo, 'A mortalidade baixou de 6% para 3%' ao invés de 'A mortalidade baixou 50%'.
8. No caso de ensaios randomizados, a análise é identificada como *intent-to-treat* ou análise de paciente avaliável.
9. Para levantamentos, o índice de respostas é fornecido em Resultados ou Delineamento.
10. Para análise multivariada, os fatores controlados no modelo são mencionados de modo bem sucinto.
11. As conclusões resultam de informações contidas no resumo.

Esses critérios são agora adotados para rever e corrigir resumos. Winkler relatou uma melhoria impressionante da qualidade após a implantação desse programa, e Pitkin et al. (2000), em pesquisa independente, também constataram notável melhoria, embora não no nível impressionante citado por Winkler. Anteriormente, Pitkin e Branagan (1998) relataram, como resultado de um ensaio randomizado controlado, que instruções específicas dadas aos autores que estavam revisando seus manuscritos não foram eficazes para diminuir as deficiências dos resumos. Parece que enviar aos autores instruções sobre a qualidade dos resumos não é, em si, garantia de melhoria, embora tais instruções sejam eficazes quando usadas pelos editores de revistas na avaliação dos resumos.

Hartley (2000a) comparou a exatidão de resumos estruturados com a de resumos 'tradicionais' de um mesmo grupo de artigos submetidos para publicação nas revistas da British Psychological Society. Isso foi possível porque os autores haviam enviado resumos tradicionais ao submeter os originais, porém,

depois que os trabalhos foram aceitos para publicação, foi-lhes exigido que apresentassem versões estruturadas. Hartley relata poucas inexatidões em qualquer um dos tipos de resumos, e que os estruturados não eram melhores nem piores do que os outros. Esse último resultado talvez não surpreenda muito, pois a maioria dos autores simplesmente converteu o resumo original para a forma estruturada. Mais difícil de explicar é esses resumos de psicologia parecerem mais exatos do que os resumos de medicina dos estudos de Pitkin.

O valor de previsibilidade dos resumos (isto é, sua capacidade de indicar a relevância do item de que deriva para os interesses de algum usuário) é examinado no próximo capítulo.

Questões de compatibilidade

Há 50 anos, a única razão existente para que fossem redigidos resumos era a de criar a representação de um documento que seria lida por seres humanos. Entretanto, os resumos são hoje utilizados com uma segunda finalidade: proporcionar uma representação que sirva para buscas feitas por computador. Infelizmente, essas duas finalidades não são inteiramente compatíveis. Para os objetivos da recuperação, a redundância é conveniente. Quer dizer, um tópico estará mais bem representado se o for de várias formas. Por exemplo, a inclusão dos sinônimos 'asas de vôo livre' e 'asas deltas' em alguns resumos aumenta a probabilidade de o item ser recuperado — um consulente usará 'vôo livre' e o outro poderá pensar em 'asa delta'. Para o leitor humano, por outro lado, é melhor haver coerência do que redundância. Na realidade, o usuário se sentirá muito confuso se as mesmas idéias forem descritas de diferentes formas no resumo.

Para os objetivos da recuperação, quanto mais longo for o resumo melhor será. Pelo menos, quanto mais longo for o resumo mais pontos de acesso proporcionará, e quanto mais pontos de acesso houver maior será o potencial de alta revocação na recuperação. Ao mesmo tempo, temos de admitir que provavelmente haverá perda de precisão: quanto mais extenso for o resumo mais aspectos 'secundários' do documento serão introduzidos e maior será o potencial de falsas associações (ver capítulos 6, 11 e 14). Para o leitor humano, a brevidade é certamente conveniente. Ela também convém para os assinantes de serviços impressos, pois resumos mais longos geram publicações mais caras.

Para o leitor humano, é útil a menção de aspectos negativos: por exemplo, 'porém exclui considerações sobre custos' informa ao leitor sobre o que não deve esperar encontrar no documento. A inclusão da palavra 'custos' no resumo fará, evidentemente, com que ele seja recuperado em buscas nas quais o custo seja um aspecto importante — exatamente a situação na qual esse resumo não deveria ser recuperado.

Para os objetivos da recuperação, também é melhor evitar certas palavras ou expressões. A locução comum 'lançar mão de' criará problemas em muitas bases de dados, pois levará à recuperação de itens sobre a parte do corpo humano

— mão —, e a flexão verbal 'cobre', do verbo 'cobrir', fará recuperar itens sobre o metal *cobre*. Portanto, para uma recuperação mais eficaz, os resumidores devem evitar termos que sabidamente causarão problemas desse tipo.

Até mesmo as convenções relativas a pontuação e sintaxe, que têm sentido para o leitor humano, podem criar problemas para o computador. Suponhamos, por exemplo, uma frase que termine com a palavra 'precipitação' seguida imediatamente por outra que comece com a palavra 'ácidos'. Em muitos sistemas este item será recuperado durante uma pesquisa sobre 'precipitação de ácidos', embora nada tenha a ver com este assunto.

Os minirresumos de Lunin (1967) (ver capítulo anterior), ao contrário do resumo convencional, destinam-se basicamente a facilitar as buscas por computador. Embora possam ser interpretados por usuários inteligentes, são definitivamente mais difíceis de ler e entender, e se ignora como um enunciado esquemático como esse seria aceito pelos usuários de um sistema de recuperação.

Tudo isso aponta para o fato de que um resumo 'ideal' para o leitor pode não ser ideal para as buscas informatizadas. Mas, até onde se pode prever, os resumos continuarão a servir a ambas as finalidades. Mesmo que a importância dos serviços impressos venha a declinar, os resumos ainda serão necessários como um produto intermediário em buscas informatizadas. Uma das implicações disso é que as editoras de serviços secundários terão de rever suas instruções, para que os resumidores passem a criar resumos que, na medida do possível, sejam sucedâneos eficazes tanto para a realização de buscas quanto para a leitura.

Fidel (1986) prestou um grande serviço ao analisar as instruções para redação de resumos de 36 produtores de bases de dados. A síntese que ela fez das instruções que parecem ser relevantes para as características de recuperabilidade dos resumos é reproduzida na figura 52. Mais que tudo, sua síntese revela algumas divergências de opinião: utilize a linguagem do autor, não utilize a linguagem do autor; utilize linguagem idêntica à linguagem dos termos de indexação atribuídos, utilize linguagem que complemente os termos atribuídos, e assim por diante. A regra mais sensata talvez seja a que especifica que o resumo deve incluir termos relevantes que faltem nos descritores e no título. Muitas vezes, esses serão termos mais específicos do que os do vocabulário controlado.

Booth e O'Rourke (1997) estudaram resumos estruturados de medicina num contexto de recuperação da informação. Por meio da importação de registros do MEDLINE, conseguiram criar duas bases de dados em que podiam fazer buscas, sendo uma de resumos completos e a outra de resumos segmentados em vários componentes (objetivos, delineamento, conclusões, e assim por diante) da estrutura. As buscas feitas na base de dados segmentada, naturalmente, obtiveram maior precisão, porém menor revocação. Os consulentes também tiveram dificuldade para decidir em quais segmentos fazer as buscas.

Nomoto e Matsumoto (2001) defendem a avaliação da qualidade de resumos produzidos automaticamente (na realidade, extratos) em termos de quão satisfatória seria a possibilidade de substituir os textos integrais nas tarefas de

recuperação da informação. Eles parecem acreditar que esta idéia se originou com eles, quando, de fato, é bastante antiga.

O boletim interno

O fato de haver bases de dados bibliográficos em praticamente todos os campos do conhecimento e de em alguns deles haver várias bases de dados concorrentes não elimina inevitavelmente a necessidade de um boletim de resumos destinado à clientela interna de uma instituição. O centro de informações de uma empresa ou outro tipo de organização em que haja um forte programa de pesquisas pode almejar produzir seu próprio boletim em virtude de:

1. Os periódicos de resumos existentes não serem suficientemente atuais na cobertura de materiais fundamentais e do maior interesse para a instituição.
2. Nenhuma base de dados, isoladamente, em formato impresso ou eletrônico, abranger, provavelmente, todos os materiais de interesse para a instituição. Na realidade, muitas bases de dados são relevantes para os interesses da instituição quando se tem em conta a diversidade de conteúdo temático e de formas documentais.
3. Nenhuma base de dados externa abrangerá certos materiais de importância, e, de modo mais evidente, os relatórios internos da própria instituição, literatura de fabricantes, material publicitário dos concorrentes, etc.

Para otimizar os procedimentos empregados na produção do boletim interno, será preciso identificar os materiais que serão resumidos diretamente. Estes certamente incluirão os relatórios internos da própria empresa e materiais externos considerados de especial importância. Por exemplo, alguém pertencente ao quadro de pessoal do centro poderá examinar todas as patentes novas e preparar resumos daquelas que se revistam de possível interesse para a empresa — o que é, em si mesmo, uma arte. Valendo-se dos métodos a serem examinados no capítulo seguinte, será identificada uma 'lista básica' de periódicos que, quase com certeza, são extraordinariamente produtivos no que concerne aos interesses da instituição. Esses periódicos também serão resumidos diretamente.

É possível que as fontes analisadas dessa forma regularmente produzam, por hipótese, de 80 a 90% da bibliografia a ser incluída no boletim interno. Para elevar essa cobertura bem acima do nível de 90% será preciso utilizar fontes impressas de caráter mais genérico. Os membros da equipe que analisa os periódicos pertencentes à lista básica à procura de artigos de interesse devem também examinar os serviços de indexação/resumos em formato impresso que forem apropriados. Isto revelará outros itens relevantes, como, por exemplo, os que aparecem em fontes que não são adquiridas por assinatura diretamente. Uma fonte abrangente no campo científico, como o *Chemical Abstracts*, é particularmente útil para a localização de itens de interesse potencial.

O Conteúdo dos Resumos

Enunciados gerais

Empregue conceitos e termos 'importantes' (p. ex., aqueles que melhorarão a recuperação em texto livre; aqueles sobre os quais o documento contém bastante informação; ou palavras-chave).

Termos de indexação

Coordene os conceitos usados nos resumos com os descritores atribuídos.

a) Inclua nos resumos conceitos que sejam idênticos aos descritores.
b) Inclua nos resumos conceitos que complementem os descritores (p. ex., termos relevantes que faltem na indexação com descritores e nos títulos, termos mais específicos do que os descritores, ou determinado tipo de termo importante para a área de assunto, como nomes geográficos).
c) Inclua nos resumos conceitos que complementem ou sejam idênticos aos descritores.

Contribua para melhorar a indexação independentemente da linguagem de indexação utilizada.

Listas de conferência

Obedeça a uma lista de elementos relacionados à recuperação que serão incluídos nos resumos.

Formas de listas de conferência:

a) Categorias que serão incluídas nos resumos (p. ex., materiais, propriedades e processos) e as condições que determinarão sua inclusão (p. ex., somente quando forem analisadas detidamente, ou sempre que forem mencionadas).
b) Diretrizes específicas e determinadas (p. ex., 'sempre que tratar de um novo produto, mencione o nome da empresa').

A Linguagem dos Resumos

Emprego da linguagem do autor

Empregue a linguagem do autor.

Não empregue a linguagem do autor.

a) Empregue termos correntes e determinados, específicos da área temática.

Empregue tanto a linguagem do autor quanto sinônimos.

Relação com a linguagem de indexação utilizada

Coordene os termos nos resumos com os descritores.

Complemente os descritores com termos nos resumos (p. ex., empregue sinônimos ou termos mais específicos).

Empregue termos específicos e de uso reconhecido para categorias determinadas (tais como materiais, processos e produtos).

Práticas a evitar

Não empregue a negativa (p. ex., use doente ao invés de que *não goza saúde*).

Não use termos em forma de lista que tenha uma última palavra em comum como se fosse uma série (tais como 'pequenos, médios e grandes países').

Formas das palavras

Adote as práticas linguísticas locais (p. ex., mude a ortografia norte-americana quando se tratar de bases de dados inglesas).

Expresse sempre pormenorizadamente os termos de certas categorias (p. ex., processos, materiais, produtos).

Quando um termo e um descritor forem iguais, registre o termo na forma adotada pelo descritor.

Expresse os termos tanto em sua forma abreviada quanto em sua forma por extenso.

FIGURA 52
Regras, destinadas a resumidores, concernentes às características de recuperabilidade dos resumos
Reproduzidas de Fidel (1986) com permissão de Emerald

Pode-se perguntar por que, em 2003, alguém consultaria serviços secundários impressos ao invés de regularmente fazer buscas em linha nas bases de dados apropriadas. Este seria o modo de atuação preferido de uma instituição cujos interesses estivessem claramente delimitados e que pudessem ser expressos de forma bastante abrangente numa estratégia de busca. Algumas organizações, porém, têm tal diversidade de interesses heterogêneos que se torna muito difícil localizar itens de interesse potencial, salvo mediante consulta a amplas seções de fontes publicadas. Ademais, a serendipidade desempenha aqui importante papel: um bom especialista em informação pode identificar itens relevantes para uma empresa que talvez estejam fora de seu perfil de interesse, como, por exemplo, uma nova aplicação potencial para um produto da empresa.

De qualquer modo, o boletim interno será compilado mediante a análise tanto de fontes primárias quanto secundárias, estas complementando a cobertura das anteriores. Num grande centro de informação, a equipe responsável pela análise da literatura incluiria algumas pessoas que teriam como tarefa principal o exame de materiais estrangeiros, a redação de resumos no vernáculo e a realização de traduções integrais de itens julgados bastante importantes.

Quanto à redação mesma dos resumos, as pessoas incumbidas disso economizarão muito tempo ao fazerem marcações no texto do próprio documento, a fim de que a entrada de dados seja feita diretamente da publicação. Em alguns casos será possível utilizar diretamente os resumos de autor, ou necessitarão de alguma alteração, como cortes ou acréscimos. Em outros casos, pode-se elaborar um 'resumo' perfeitamente satisfatório extraindo-se porções do texto, talvez da parte correspondente às conclusões ou resultados. Naturalmente, sempre haverá alguns itens que exigirão a redação de resumos originais, seja porque não exista um resumo satisfatório, seja porque o processo de elaboração do extrato é inadequado, ou porque algum aspecto de grande interesse para a empresa, porém de interesse secundário para o autor, precisa ser ressaltado.

Os resumos preparados para uso interno podem ser disseminados de vários modos. Destes, o mais comum é um boletim duplicado mecanicamente e que seja editado com regularidade. Tendo em vista que o mesmo pode ser considerado como um instrumento de informação da maior importância para a empresa, deveria, se possível, ser editado semanalmente. Os resumos seriam organizados em seções que permaneceriam mais ou menos constantes, ao longo das semanas, de modo a facilitar a consulta. Seria incluído um sumário analítico, com indicação de seções e subseções. Um boletim desse tipo pode conter de 80 a 150 resumos. A cada resumo é atribuído um número exclusivo para fins de identificação e ordenação. Deve haver um formulário apenso ao boletim para que seus destinatários encaminhem pedidos dos documentos resumidos.

O boletim de resumos será distribuído para os nomes constantes de uma lista de destinatários. Para certos nomes-chave da organização, o centro de informação poderá fazer algo mais, afixando um memorando à capa do boletim, que chamará a atenção de cada uma dessas pessoas para itens que talvez sejam especialmente relevantes. A forma convencional de expressar isso seria mais ou menos a seguinte: 'Se seu tempo só for suficiente para examinar poucos itens, é provável que os seguintes sejam de seu particular interesse.'

Uma alternativa ao boletim como tal é, evidentemente, disseminar os resumos como itens separados. Isso requer que os disseminadores possuam uma imagem nítida e abrangente dos interesses individuais, de modo que cada pessoa receba somente itens que lhe sejam potencialmente pertinentes, ou que algum programa de computador seja utilizado para cotejar características dos resumos com perfis de interesses individuais.

Realmente não é recomendável a distribuição de resumos separados. Isso exige muito mais trabalho de parte do centro de informação e elimina a possibilidade de o usuário encontrar outras informações percorrendo as páginas a esmo. Um boletim bem-organizado é um instrumento de disseminação mais eficaz. Chamar a atenção para itens selecionados do boletim, com o objetivo de poupar tempo a pessoas-chave, é um substituto eficaz da disseminação de resumos separados.

Ao criar um boletim interno, o centro de informação estará, evidentemente, formando uma base de dados. Além disso, trata-se de uma base de dados que será de grande utilidade potencial para a instituição. Deverá ser acessível em linha dentro da empresa, de uma forma que se preste a buscas eficazes. Cada resumo pode ser indexado (pela própria pessoa que o redige), seu texto prestar-se a buscas ou o sistema de recuperação adotar uma combinação de termos de indexação com expressões do texto.

É claro que a intranet da própria instituição pode ser usada para disseminar resumos eletronicamente para as pessoas e/ou tornar o boletim acessível na íntegra para consultas em linha. Não obstante, ainda há muitos argumentos favoráveis à utilidade para consulta de um boletim distribuído em formato impresso.

Inclinação para um assunto

Mencionou-se a inclinação para um assunto no capítulo anterior. Quando uma publicação de resumos é projetada para ser utilizada por um grupo de pessoas que possuem interesses claramente definidos e especializados (como seria o caso de um boletim interno), é conveniente, sem dúvida, que cada resumo seja moldado aos interesses precisos do grupo. Isso foi reconhecido no projeto de análises de conteúdo modulares (Lancaster et al., 1965) descrito no capítulo 7. Para que essas análises tivessem o máximo de utilidade para um grupo diversificado de serviços secundários, propôs-se que incorporassem 'módulos temáticos'. Uma análise de conteúdo incluiria um resumo 'básico' mais parágrafos suplementares, cada um dos quais seria moldado aos interesses de determinado grupo. As entradas de índice fornecidas também refletiriam essa diversidade de interesses. O apêndice 2 exemplifica o método: o resumo básico sobre contato de chama é complementado com parágrafos que relacionam o trabalho a interesses em fisiologia e medicina, à indústria de plásticos, à indústria da borracha e às indústrias de roupas de proteção e aeronáutica.

CAPÍTULO 9

Aspectos da avaliação

O tema da avaliação é tratado em diversos capítulos deste livro. O capítulo 1, por exemplo, refere-se aos critérios de avaliação dos resultados de buscas realizadas numa base de dados, enquanto o capítulo 6 focaliza a qualidade da indexação e os critérios segundo os quais essa qualidade pode ser aferida.

A indexação e a redação de resumos não são atividades que devam ser consideradas como fins em si mesmas. São os resultados dessas atividades que devem ser avaliados e isso somente pode ser feito no contexto de determinada base de dados, seja ela em formato impresso ou eletrônico. Nesse contexto, a indexação é avaliada como bem-sucedida quando permite a quem realiza as buscas localizar itens de que precisa sem ter de examinar muitos de que não precisa. Os resumos são bem-sucedidos quando permitem prever corretamente quais os documentos que serão úteis a um consulente e quais não serão, ou se são úteis como substitutos do documento em buscas textuais.

Uma base de dados bibliográficos não pode ser avaliada de forma isolada, mas somente em função de sua utilidade ao responder a várias necessidades de informações. No que concerne a determinada necessidade de informação, avalia-se uma base de dados de acordo com quatro critérios principais:

1. *Cobertura*. Quantos documentos sobre um assunto, publicados durante determinado período, se acham incluídos na base de dados?
2. *Recuperabilidade*. Quantos documentos sobre o assunto, incluídos na base de dados, são encontrados com o emprego de estratégias de busca 'razoáveis'?
3. *Previsibilidade*. Ao utilizar informações da base de dados, com que eficiência o usuário pode aferir quais os itens que serão e os que não serão úteis?
4. *Atualidade*. Os itens publicados recentemente são recuperáveis, ou atrasos na indexação/redação de resumos provocam uma situação em que os itens recuperados mostram resultados de pesquisas 'antigos' ao invés de 'novos'?

Cobertura

A avaliação da *cobertura* de uma base de dados é bastante semelhante à avaliação da completeza do acervo de uma biblioteca em relação a um assunto. Na realidade, o acervo de livros de uma biblioteca é em si mesmo uma base de dados, do mesmo modo que o catálogo da biblioteca — um é uma base de dados de artefatos, e o outro, uma base de dados de representações desses artefatos.

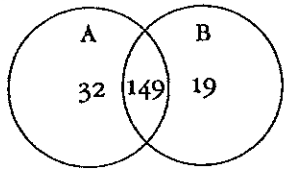
Uma forma de avaliar a cobertura do acervo de uma biblioteca sobre determinado assunto consiste em obter bibliografias confiáveis sobre esse assunto e cotejá-las com o acervo. Esta técnica também pode ser aplicada à avaliação da cobertura de serviços de indexação/resumos. Martyn (1967) e Martyn e Slater (1964) exemplificaram o emprego desse método. Suponhamos, por exemplo, que se queira avaliar a cobertura do *Index Medicus* sobre o assunto leucemia felina. Com sorte, poder-se-á encontrar uma bibliografia que pareça ou afirme ser exaustiva a respeito desse assunto para determinado período. Neste caso, a tarefa é fácil: cotejam-se as entradas da bibliografia com o índice de autores do *Index Medicus*, a fim de determinar quais os itens que são e os que não são incluídos. Como resultado disso conclui-se que o *Index Medicus* cobre, digamos, 84% mais ou menos da literatura sobre esse assunto. Evidentemente, é preciso conhecer algo a respeito das diretrizes adotadas pela base de dados que está sendo avaliada; por exemplo, que o *Index Medicus* se dedica quase exclusivamente a artigos de periódicos e não inclui monografias.

Essa técnica não está isenta de problemas. Em primeiro lugar, não é fácil encontrar bibliografias exaustivas. Além disso, talvez se ignore totalmente como uma bibliografia foi compilada. Se a bibliografia sobre leucemia felina tiver sido compilada basicamente com a utilização do *Index Medicus* (ou seu equivalente eletrônico) sua utilidade será muito limitada para a avaliação desta ferramenta.

O fato é que, evidentemente, não precisamos realmente de uma bibliografia exaustiva para avaliar a cobertura de uma base de dados sobre um assunto; basta uma amostra de itens que seja representativa. Uma forma de obter essa amostra é usar uma base de dados como fonte de itens com os quais será avaliada a cobertura de outra base de dados. Suponhamos, por exemplo, que alguém quisesse saber o grau de completeza da cobertura do *Engineering Index* em relação ao assunto supercondutores. Consultar-se-ia o *Physics Abstracts*, a fim de identificar, por hipótese, 200 itens que este serviço tivesse indexado sob 'supercondutores' ou 'supercondutividade', e este conjunto seria utilizado para calcular a cobertura do *Engineering Index*. Depois de cotejá-lo com os índices de autores do *Engineering Index,* verificar-se-ia que ali se encontram 142/200, o que nos dá uma cobertura estimativa de 71%. O fato de os 200 itens não serem *todos* os itens publicados sobre supercondutores não é importante; trata-se, em certo sentido, de um conjunto 'representativo' de itens sobre supercondutores e constitui uma amostra perfeitamente legítima para se usar na estimativa de cobertura.

Seria possível, obviamente, fazer o percurso inverso, empregando itens extraídos do *Engineering Index* para avaliar a cobertura do *Physics Abstracts.* Dessa forma também se determina o grau de duplicidade e exclusividade de dois (ou mais) serviços, como se acha representado no diagrama no final deste parágrafo. Obtêm-se esses resultados extraindo-se de *A* uma amostra aleatória de itens sobre supercondutores e cotejando-os com *B*, e extraindo-se de *B* uma amostra aleatória de itens sobre supercondutores e cotejando-os com *A*. Essas

amostras nos permitiriam estimar a cobertura de *A* (181/200 ou cerca de 90% no exemplo hipotético acima), a cobertura de *B* (168/200 ou cerca de 84%), o grau de duplicidade entre os serviços (149/200 ou cerca de 75%), e a exclusividade (cerca de 16% dos itens incluídos por *A*, ou seja, 32/200, aparecem exclusivamente nesse serviço enquanto a cifra comparável para *B* está um pouco abaixo de 10% (19/200)). O mesmo tipo de resultado seria alcançado, e sob certos aspectos mais facilmente, se extraíssemos uma amostra de uma terceira fonte, *C*, para estimar a cobertura, a duplicidade e a exclusividade de *A* e *B*.



Nas considerações acima, pressupôs-se a avaliação de uma base de dados em formato impresso. Os procedimentos não difeririam de modo significativo caso fossem aplicados a uma base de dados em formato eletrônico. É maçante, sem dúvida, dar entrada a talvez centenas de nomes de autores, a fim de determinar a cobertura de uma fonte em linha. A solução deste problema está em realizar, inicialmente, uma ampla busca por assunto (de qualquer modo necessária, se se quiser determinar a recuperabilidade; ver comentários adiante), e, em seguida, fazer buscas complementares por autor. Adotando o mesmo exemplo, extrair-se-ia uma amostra de itens indexados sob SUPERCONDUTORES ou SUPERCONDUTIVIDADE da base de dados INSPEC, a fim de avaliar a cobertura deste assunto no COMPENDEX. O primeiro passo seria fazer uma busca no COMPENDEX sob os termos relativos a supercondutores, a fim de verificar quantos dos itens da amostra teriam sido recuperados. O passo seguinte seria realizar buscas por autor, a fim de determinar se os itens da outra amostra apareciam ou não no COMPENDEX e, em caso positivo, descobrir como foram indexados.

Existe a possibilidade de ocorrer um problema quando se trata de extrair uma amostra de itens de uma base de dados para avaliar a cobertura de outra. Em alguns casos uma base de dados em formato impresso indexará os itens apenas sob os termos considerados 'mais importantes'. Isso acontece com o *Index Medicus,* por exemplo, de modo que itens indexados sob o termo FELINE LEUKEMIA VIRUS [vírus da leucemia felina] serão somente aqueles que tratam do assunto de modo predominante e não os que tratam do mesmo assunto de modo periférico. Ao utilizar uma amostra extraída do *Index Medicus* para avaliar outro serviço, temos, portanto, de admitir que a estimativa da cobertura desse serviço diz respeito apenas à cobertura de artigos de periódicos que tratam 'predominantemente' do assunto. No entanto, se extrairmos nossa amostra da base de dados MEDLINE (fundamentalmente o equivalente eletrônico do *Index Medicus*), não teremos esse problema, pois um termo de indexação como FELINE LEUKEMIA

VIRUS será ali empregado para se referir a este assunto quando abordado de modo periférico, bem como quando abordado de modo predominante. Também em certos índices impressos não é feita qualquer distinção entre termos 'mais importantes' e 'menos importantes'. Por exemplo, uma amostra de assunto poderia ser extraída de um dos índices da *Excerpta Medica* com a expectativa razoável de que os itens escolhidos incluiriam alguns em que o assunto é tratado de maneira que não chega a ser predominante.

Obviamente, ao extrair amostras de um serviço de indexação/resumos para avaliar outro, levam-se em conta as datas de publicação. Por exemplo, pode-se extrair uma amostra de itens incluídos na *Excerpta Medica* durante o ano de 1997. Se for utilizada para avaliar a cobertura do *Index Medicus*, provavelmente serão confrontados em primeiro lugar os índices de autores correspondentes a 1997. Quaisquer itens que não forem aí encontrados serão cotejados com os índices de 1998 (e talvez até posteriores) ou 1996 (e mesmo, em alguns casos, anteriores), tendo em vista que a National Library of Medicine não terá necessariamente indexado os documentos na mesma época em que o fez a Excerpta Medica Foundation. Ao agir assim, pode-se, evidentemente, ter alguma idéia da atualidade relativa das duas ferramentas. Mais adiante, neste capítulo, trataremos da questão da atualidade.

Há outra fonte que pode ser utilizada para avaliar a cobertura de uma base de dados: as referências bibliográficas que aparecem nos artigos de periódicos. Voltando ao exemplo já citado, suponhamos que identificamos uma quantidade de artigos publicados recentemente em periódicos científicos que tratam de leucemia felina. As referências bibliográficas incluídas nesses artigos serão usadas para compor uma bibliografia a ser aplicada na avaliação da cobertura do *Index Medicus* ou de um dos índices da *Excerpta Medica.*

Há uma evidente diferença entre utilizar itens retirados de bibliografias sobre leucemia felina (ou itens indexados sob esse termo em alguma ferramenta bibliográfica) e utilizar referências bibliográficas de artigos de periódicos: os primeiros, provavelmente, serão itens que tratam de leucemia felina de per si enquanto os últimos são as fontes de que necessitam os pesquisadores que atuam na área da leucemia felina. É provável que estas últimas fontes ultrapassem bastante o assunto específico e, de fato, abranjam um amplo setor das ciências biológicas e talvez até de outros campos. O avaliador pode optar por excluir quaisquer itens que lhe pareçam periféricos ao tópico da avaliação ou incluí-los, com a justificativa de que uma ferramenta bibliográfica, para que seja útil para o pesquisador desse assunto, deve proporcionar acesso a todos os materiais afins necessários à fundamentação de sua pesquisa.

Na avaliação de uma base de dados que se restrinja quase exclusivamente a artigos de periódicos (como é o caso do *Index Medicus*), poder-se-ia tomar um atalho óbvio para chegar a uma estimativa de cobertura. Tendo extraído uma amostra de outra fonte, ou fontes, identificar-se-iam os artigos de periódicos e

depois simplesmente se faria uma checagem para verificar se esses periódicos são regularmente abrangidos pelo *Index Medicus*. Com toda a probabilidade isso daria uma estimativa de cobertura razoável. Se se quisesse ser mais preciso, entretanto, os itens da amostra (ou pelo menos um subconjunto extraído aleatoriamente) seriam checados por nome de autor, devido ao fato de certos periódicos serem indexados apenas seletivamente, e de alguns artigos (e talvez fascículos completos de alguns periódicos) que deveriam ter sido indexados não o serem por algum motivo.* O atalho que passa pelos títulos dos periódicos é menos útil para a avaliação da cobertura de uma base de dados que inclua itens publicados de todos os tipos, e não tem utilidade alguma no caso de uma base de dados altamente especializada que procure incluir tudo sobre determinado assunto, de qualquer fonte, sem se restringir a determinado conjunto de periódicos.

Há várias razões possíveis que justificam uma avaliação de cobertura. Por exemplo, um centro de informação quer saber se determinada base de dados, como a do *Chemical Abstracts*, cobre de forma exaustiva sua área de especialização ou se precisaria recorrer a várias bases de dados para conseguir cobertura mais completa. Também o produtor de uma base de dados pode estar interessado em saber em que medida ela cobre satisfatoriamente determinada área. Neste caso, seria importante determinar quais os tipos de publicações que oferecem maior cobertura e os que oferecem menor cobertura. Para tanto, seria preciso classificar os itens abrangidos e os não abrangidos, segundo certas características, como tipo de documento, língua, lugar de publicação e título do periódico.

A partir desses dados poder-se-ia determinar como seria possível melhorar a cobertura de modo a proporcionar a melhor relação custo–eficácia. Ao estudar a cobertura de bases de dados é importante estar atento ao fenômeno da *dispersão*. Este fenômeno prejudica as bases de dados altamente especializadas, bem como a biblioteca ou centro de informação muito especializado, e favorece a base de dados, biblioteca ou centro de teor mais geral. Vejamos, por exemplo, um centro de informação sobre AIDS, cuja meta seja colecionar a bibliografia desse assunto de modo exaustivo e assim criar uma base de dados abrangente. As dimensões deste problema são exemplificadas nas figuras 53–59 que se baseiam em buscas feitas na base MEDLINE em 1988. A figura 53 mostra que somente 24 artigos de periódicos sobre AIDS foram publicados até o final de 1982; no ano de 1987 esta bibliografia alcançou 8 510 itens. Em 1982, toda a bibliografia de AIDS se limitava a três idiomas, porém, em 1987, eram 25 as línguas utilizadas e 54 os países que contribuíam para essa literatura (figuras 54 e 55). Mais eloqüente é a figura 56, que mostra que toda a bibliografia de AIDS se achava em apenas 14 periódicos em 1982, mas em 1987 a participação era de quase 1 200 periódicos!

---

* Por exemplo, Thorpe (1974), ao estudar a literatura de reumatologia, obteve uma estimativa de cobertura para o *Index Medicus* com base nos títulos de periódicos que foi um tanto diferente daquela baseada nos artigos dos periódicos. Brittain e Roberts (1980) também apresentam indicações sobre a necessidade de estudar a cobertura e a duplicidade no âmbito dos artigos.

Todos esses exemplos demonstram o fenômeno da *dispersão*. À medida que cresce, a bibliografia de um assunto torna-se cada vez mais dispersa (mais países presentes, mais línguas utilizadas, mais periódicos que publicam, maior variedade de documentos) e, portanto, mais difícil de identificar, coletar e organizar.

Ano	*Número de itens publicados*	*Total acumulado de publicações*
1982	24	24
1983	641	665
1984	1 158	1 823
1985	1 707	3 530
1986	2 117	5 647
1987	2 863	8 510

FIGURA 53
Crescimento da literatura científica sobre AIDS, 1982–1987 (Fonte: MEDLINE)

	1982	*1983*	*1984*	*1985*	*1986*	*1987*
Número de idiomas	3	14	21	21	20	23
Número acumulado de idiomas	3	14	22	25	25	25

FIGURA 54
Literatura sobre AIDS: cobertura por idioma, 1982–1987 (Fonte: MEDLINE)

	1982	*1983*	*1984*	*1985*	*1986*	*1987*
Número de países produtores	5	30	38	43	39	42
Número acumulado de países produtores	5	30	39	48	52	54

FIGURA 55
Literatura sobre AIDS: cobertura por país, 1982–1987 (Fonte: MEDLINE)

Ano	*Número de periódicos*	*Número acumulado de periódicos*
1982	14	14
1983	228	234
1984	257	464
1985	492	719
1986	582	952
1987	676	1 170

FIGURA 56
Número de periódicos que publicaram artigos sobre AIDS, 1982–1987 (Fonte: MEDLINE)

N.º de periódicos	*N.º de artigos*	*N.º acumulado de periódicos*	*N.º acumulado de artigos*	*N.º de periódicos*	*N.º de artigos*	*N.º acumulado de periódicos*	*N.º acumulado de artigos*
1	550	1	550	2	29	42	3 954
1	351	2	901	3	28	45	4 038
1	307	3	1 208	5	27	50	4 173
1	303	4	1 511	2	26	52	4 225
1	289	5	1 800	7	25	59	4 400
1	217	6	2 017	3	24	62	4 472
1	200	7	2 217	3	23	65	4 541
1	104	8	2 321	3	22	68	4 607
1	98	9	2 419	2	21	70	4 649
1	97	10	2 516	5	20	75	4 749
1	83	11	2 599	4	19	79	4 825
1	78	12	2 677	7	18	86	4 951
1	70	13	2 747	7	17	93	5 070
2	67	15	2 881	4	16	97	5 134
1	60	16	2 941	7	15	104	5 239
1	59	17	3 000	8	14	112	5 351
1	54	18	3 054	14	13	126	5 533
1	52	19	3 106	12	12	138	5 677
1	49	20	3 155	13	11	151	5 820
1	48	21	3 203	11	10	162	5 930
2	47	23	3 297	15	9	177	6 065
2	46	25	3 389	14	8	194	6 101
2	40	27	3 469	40	7	234	6 481
1	39	28	3 508	42	6	276	6 733
1	36	29	3 544	50	5	326	6 983
2	34	31	3 612	87	4	413	7 331
4	33	35	3 744	117	3	530	7 682
1	32	36	3 776	188	2	718	8 058
4	30	40	3 896	452	1	1 170	8 510

FIGURA 57
Dispersão da literatura de periódicos sobre AIDS em 1987 (Fonte: MEDLINE)

O aspecto mais impressionante da dispersão diz respeito à separação de artigos entre os títulos de periódicos. Foi Bradford quem primeiro observou este fenômeno, em 1934, fenômeno ao qual nos referimos atualmente como Lei da Dispersão de Bradford. Ela está demonstrada nitidamente na figura 57, que apresenta a dispersão de artigos de periódicos sobre AIDS no período 1982–1987. O primeiro periódico da lista participou com 550 trabalhos num período de seis anos, o segundo com 351 trabalhos e o terceiro com 307 trabalhos.

Observe-se que dois periódicos contribuíram com 67 trabalhos cada um, dois com 47 cada um, e assim sucessivamente até o fim da lista, onde temos 452

periódicos que participaram com apenas um único artigo cada um para a bibliografia de AIDS durante seis anos. Bem mais de um terço da literatura acha-se concentrado em apenas 15 periódicos. Para alcançar o terço seguinte, é preciso, no entanto, acrescentar mais 123 periódicos, enquanto o terço final acha-se disperso em mais de mil periódicos adicionais. Esta distribuição proporciona uma demonstração eloqüente da lei dos rendimentos decrescentes. Isso é revelado de modo ainda mais nítido na figura 58, que representa graficamente a percentagem de artigos em comparação com a percentagem de periódicos que contribuíram com artigos. Note-se que, à medida que se ascende na curva, a dispersão de artigos entre os títulos de periódicos cresce em proporção aproximadamente geométrica: o primeiro terço dos artigos em 15 periódicos, o segundo em 123 periódicos (15 × 8,2), e o terço final em 1 008 periódicos (numa aproximação grosseira, 15 × $8{,}2^2$). Esta é uma distribuição tipicamente bradfordiana.

É evidente que um centro de informação que esteja formando uma base de dados sobre o assunto AIDS não poderá montar este serviço apoiando-se na assinatura direta de todos os periódicos que publicam artigos de interesse. Contudo, a lista desses periódicos em ordem de número de artigos publicados (figura 57) pode ser utilizada para identificar uma lista básica de periódicos que mereçam ser adquiridos e examinados sistematicamente. A figura 59 mostra como seria o topo dessa lista, com base em dados de 1982–1987. Até que ponto dessa lista ordenada chegaria o centro de informação é algo que dependeria em parte de seus recursos financeiros. Entretanto, mesmo dispondo de recursos ilimitados, o centro não poderia adquirir todos os periódicos que publicam artigos sobre AIDS. Na medida em que se desce na lista ordenada, diminui a previsibilidade dos títulos dos periódicos. Assim, os dez títulos do topo em 1982–1987 talvez continuem ocupando essa posição durante os próximos cinco anos. Isso porém não é garantido. No caso da AIDS, por exemplo, existem atualmente novos periódicos dedicados exclusivamente a este assunto e que provavelmente virão a aparecer entre os dez do topo da lista durante o período de 1987 em diante, talvez até ocupando o primeiro lugar. No entanto, é bastante provável que todos os periódicos da figura 59 continuarão entre os mais produtivos sobre AIDS ainda por algum tempo. Os periódicos na faixa intermediária da distribuição (isto é, aproximadamente os do meio da tabela da figura 57) são muito menos previsíveis — poderão ou não continuar publicando artigos relacionados à AIDS. Os títulos que aparecem no pé da tabela são bastante imprevisíveis: um periódico que tenha publicado somente um artigo sobre AIDS em cinco ou seis anos talvez nunca mais venha a publicar outro artigo sobre o mesmo assunto.

Ao procurar formar uma base de dados especializada em AIDS, portanto, o centro de informação cobrirá uma parte dessa literatura por meio de assinatura direta — talvez uns 100 periódicos, mais ou menos — e identificará os outros itens que tratam de AIDS mediante buscas sistemáticas em outras bases de dados de mais amplo alcance: MEDLINE, BIOSIS, etc.

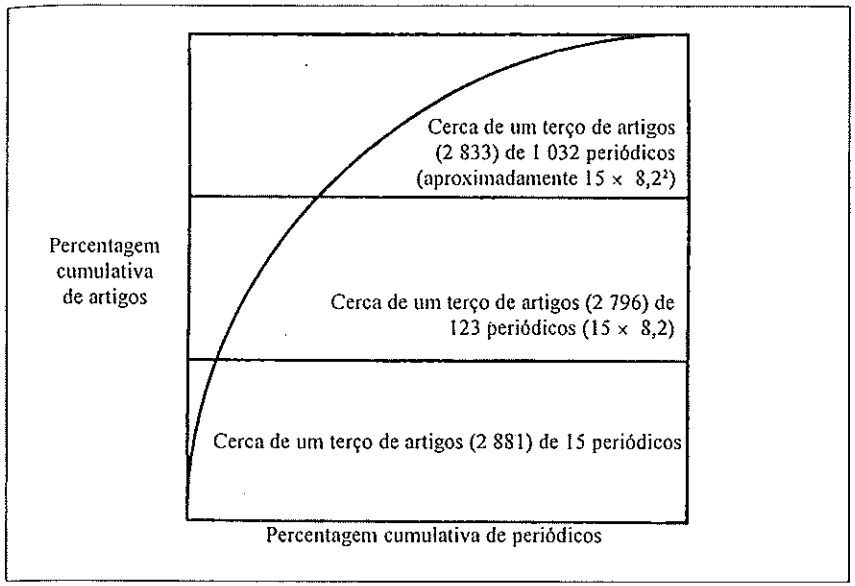


FIGURA 58
Gráfico da dispersão da literatura sobre AIDS

N.º de ordem	*Título*	*Produção*
1	*Lancet*	550
2	*Journal of the American Medical Association*	351
3	*New England Journal of Medicine*	307
4	*Annals of Internal Medicine*	303
5	*Nature*	289
6	*Science*	217
7	*British Medical Journal*	200
8	*MMWR*	104
9	*American Journal of Medicine*	98
10	*Journal of Infectious Diseases*	97

FIGURA 59
Periódicos científicos que publicaram a maioria dos artigos sobre AIDS, 1982–1987
(Fonte: MEDLINE)

Martyn (1967) e Martyn e Slater (1964) realizaram os estudos 'clássicos' sobre cobertura de serviços de indexação/resumos, porém há, na bibliografia, muitos outros trabalhos sobre cobertura ou duplicidade. Por exemplo, Goode et al. (1970) compararam a cobertura do *Epilepsy Abstracts*, um produto da Excerpta Medica Foundation, com a do *Index Medicus,* enquanto Wilkinson e Hollander (1973) compararam a cobertura do *Index Medicus* e do *Drug Literature Index*.

Dois estudos fizeram uma comparação entre *Biological Abstracts*, *Chemical Abstracts* e *Engineering Index* e seus equivalentes em formato eletrônico: Wood et al. (1972) compararam a cobertura das três fontes em termos de títulos de periódicos, enquanto Wood et al. (1973) as compararam em termos de artigos de periódicos selecionados para cobertura.

Talvez o maior estudo sobre duplicidade foi o relatado por Bearman e Kunberger (1977), que analisaram 14 serviços diferentes e quase 26 000 periódicos por eles indexados, tendo tratado da duplicidade e exclusividade de cobertura.

Embora o *Index Medicus* tenha sido analisado mais vezes do que qualquer outra fonte, a *Bibliography of Agriculture* foi tema do estudo mais intensivo sobre cobertura. Em dois relatórios afins, Bourne (1969a,b) comparou a cobertura dessa fonte com a de 15 outros serviços e calculou sua cobertura de tópicos específicos, empregando para isso as bibliografias que acompanham os capítulos de anuários de revisão da literatura.

Montgomery (1973) estudou a cobertura da literatura de toxicologia em *Chemical Abstracts*, *Biological Abstracts*, *Index Medicus*, *Excerpta Medica*, *Chemical Biological Activities* e *Science Citation Index*. Este foi um estudo inusitado, pois coletou um conjunto de 1 873 referências da literatura de toxicologia (1960–1969) junto a 221 membros da Society of Toxicology e as utilizou como base para comparação das diversas fontes.

O'Connor e Meadows (1968) estudaram a cobertura de astronomia no *Physics Abstracts*, Gilchrist (1966), a cobertura da literatura de documentação (especificamente itens sobre a avaliação de sistemas de informação) em seis serviços, e Fridman e Popova (1972), a cobertura de primatologia experimental no *Referativnyi Jurnal*. Brittain e Roberts (1980) tratam da duplicidade no campo da criminologia, e Robinson e Hu (1981) comparam a cobertura de bases de dados no campo da energia. Edwards (1976) incluiu a cobertura como um aspecto de seu estudo sobre índices em biblioteconomia e ciência da informação. La Borie et al. (1985) estudam a duplicidade em quatro serviços secundários em biblioteconomia/ciência da informação, baseando-se em títulos de periódicos, e comparam os títulos cobertos por esses serviços com aqueles cobertos por seis serviços nas ciências, inclusive ciências sociais. Outros pesquisadores estudaram a cobertura de determinados tipos de publicações (por exemplo, Hanson e Janes (1961) realizaram uma pesquisa sobre a cobertura, por parte de vários serviços, de trabalhos apresentados em eventos, e Oppenheim (1974) examinou a cobertura de patentes pelo *Chemical Abstracts*), ou a cobertura de um assunto altamente específico (por exemplo, o estudo de Smalley (1980) sobre a comparação de duas bases de dados do ponto de vista de sua cobertura da bibliografia sobre condicionamento operante).

Os estudos de cobertura são menos comuns hoje em dia, mas de vez em quando aparece algum na literatura. Brown et al. (1999), por exemplo, comparam a cobertura do *Current Index to Journals in Education* com o *Education Index*.

Estudos de cobertura ou duplicidade não são necessariamente meros exercícios intelectuais. Alguns são realizados visando a objetivos definidos, dos quais o mais evidente é como melhorar a cobertura de algum serviço. Outra finalidade desses estudos é a identificação de uma 'lista básica' de periódicos em determinado campo, identificados pelo fato de serem todos considerados merecedores de indexação por vários serviços diferentes. Um exemplo de um estudo desse tipo é relatado por Sekerak (1986), que conseguiu identificar uma lista básica de 45 periódicos no campo da psicologia a partir de um estudo sobre duplicidade entre cinco serviços da área de psicologia/atenção à saúde.

Recuperabilidade

Para quem estiver procurando informações sobre determinado assunto, será importante a cobertura de uma base de dados sobre esse assunto, principalmente se tiver de fazer uma busca exaustiva. Evidentemente, a recuperabilidade também é importante; considerando que uma base de dados inclui *n* itens sobre um assunto (o que se pode estabelecer por meio de um estudo de cobertura), quantos desses itens será possível recuperar ao fazer uma busca na base de dados?

Isso é comprovado mediante um estudo que é complementar a uma pesquisa sobre cobertura. Suponhamos que queremos estudar a cobertura e a recuperabilidade de uma variedade de assuntos que se situam no âmbito da base de dados AGRICOLA. Para cada um de dez assuntos, temos um conjunto de itens bibliográficos (estabelecido por um dos métodos antes descritos) e, para cada conjunto, sabemos quais os itens que se acham e os que não se acham incluídos no AGRICOLA. Para cada assunto teríamos uma busca realizada por um especialista em informação conhecedor do AGRICOLA, e aferiríamos a recuperabilidade com base na proporção de itens conhecidos que o especialista conseguir recuperar. Por exemplo, na primeira busca, sobre insetos daninhos à soja, sabemos que existem 80 itens sobre este tópico que se acham incluídos no AGRICOLA. O especialista, contudo, somente conseguiu encontrar 60 desses itens, ou seja, uma *revocação* (ver capítulo 1) de apenas 75%.

É claro que este tipo de estudo testa não apenas a base de dados e sua indexação, mas também a capacidade da pessoa que faz a busca. O efeito desta variável pode ser atenuado fazendo-se com que a mesma busca seja feita de modo independente por vários especialistas em informação, a fim de determinar que resultados *em média* podem ser esperados de uma busca sobre o assunto. Os resultados poderiam ser também considerados como probabilidades: por exemplo, 50/80 foram encontrados por todos os três especialistas (probabilidade de recuperação 1,00), 6/80 por dois dos três especialistas (probabilidade de recuperação 0,66), 4/80 por apenas um dos especialistas (probabilidade de recuperação 0,33), e 20/80 por nenhum deles (probabilidade de recuperação zero).

Observe-se que a recuperabilidade (revocação) é avaliada somente tendo em conta os itens conhecidos por antecipação como relevantes para o assunto da busca e que se acham incluídos na base de dados. A busca sobre pragas de insetos que atacam a soja pode recuperar um total de 200 itens, dos quais, digamos, 150 parecem relevantes. Se apenas 60 dos 80 itens 'conhecidos como relevantes' forem recuperados, a estimativa de revocação é de 0,75 o que implica que os 150 itens recuperados representam aproximadamente 75% do total de itens relevantes presentes na base de dados.

O coeficiente de revocação, evidentemente, refere-se apenas a uma dimensão da busca. A fim de estabelecer um *coeficiente de precisão* (ver capítulo 1), seria preciso que todos os itens recuperados fossem de algum modo avaliados quanto à sua relevância (por exemplo, por um grupo de especialistas no assunto). Uma alternativa seria medir a relação custo–eficácia, determinando-se o custo por item relevante recuperado. Por exemplo, o custo total de uma busca em linha (inclusive o tempo do especialista em buscas) seria de 75 dólares. Se forem recuperados 150 itens relevantes, o custo por item relevante será de 50 centavos.

Existe um modo alternativo de estudar a recuperabilidade de itens de uma base de dados, o qual envolve uma espécie de simulação. Suponhamos que sabemos existirem numa base de dados 80 itens relevantes sobre o assunto *X* e que podemos recuperar e imprimir registros que mostrem como esses itens foram indexados. Podemos, então, por assim dizer, simular uma busca registrando o número de itens recuperáveis sob vários termos ou combinações de termos. Um exemplo hipotético disso é mostrado na figura 60. Nesse caso, 38/80 itens conhecidos como relevantes para o assunto supercondutores aparecem sob o termo SUPERCONDUTORES, enquanto 12 outros são encontrados sob SUPERCONDUTIVIDADE. Não se encontram itens adicionais sob esses dois termos, mas somente sob os termos *A*–*J*. Conclui-se, a partir de uma análise desse tipo, que 50/80 itens são facilmente recuperáveis e que 62/80 seriam localizados por um especialista sagaz porque os termos *A* e *B* ou estão relacionados de perto com 'supercondutores', ou estão explicitamente ligados ao termo SUPERCONDUTORES por intermédio de remissivas na base de dados. Conclui-se ainda que 18/80 provavelmente não seriam recuperados porque aparecem somente sob termos que não têm relação direta com 'supercondutores' (por exemplo, podem representar aplicações do princípio da supercondutividade).

Albright (1979) realizou minucioso estudo desse tipo empregando o *Index Medicus*. Buscas simuladas, feitas sobre dez assuntos diferentes, revelaram que, em média, teriam de ser consultados 44 termos diferentes para recuperar todos os itens que se sabia serem relevantes para determinado assunto. Embora alguns estivessem ligados, por meio da estrutura hierárquica ou de remissivas do vocabulário do sistema, muitos não apresentavam essa ligação, e seria improvável que mesmo um especialista em buscas, persistente e habilidoso, viesse a consultá-los. A figura 61 mostra um exemplo do trabalho de Albright. Na realidade, somente um especialista em buscas que fosse muito inteligente e persistente obteria alta revocação numa busca sobre esse tópico no *Index Medicus*.

Termo	*Número de itens recuperáveis*
Supercondutores	38
Supercondutividade	12
A	7
B	5
C	3
D	3
E	3
F	2
G	2
H	2
I	2
J	1
TOTAL	80

FIGURA 60
Exemplo hipotético da distribuição de itens sobre 'supercondutores' sob termos num índice impresso

Termo *	*Número de itens recuperáveis*	*Número acumulado de itens recuperáveis*
LYMPHOCYTES	23	23
B-LYMPHOCYTES	7	30
THYMUS GLAND	6	36
CELL MEMBRANE	2	38
SWINE	2	40
ANTIGENS	1	41
ANTIBODY FORMATION	1	42
HISTOCOMPATIBILITY	1	43
GENES	1	44
ANTILYMPHOCYTE SERUM	1	45

FIGURA 61
Distribuição de itens sobre imunologia celular no porco sob termos no *Index Medicus*
Apud Albright (1979) com permissão do autor

Assim como os artigos estão dispersos pelos títulos de periódicos, os itens sobre um assunto incluídos numa base de dados estão dispersos sob muitos termos diferentes. É o que mostra graficamente a figura 62. É possível que, para

* Para a tradução destes termos, ver Descritores em Ciências da Saúde (*DeCS*) em http://decs.bvs.br/

determinado assunto, se encontre uma percentagem relativamente alta de itens relevantes sob um pequeno número de termos 'óbvios' (por exemplo, SUPERCONDUTORES ou SUPERCONDUTIVIDADE numa busca sobre supercondutores). Acrescentando outros termos bastante afins, talvez ligados aos termos na estrutura do vocabulário da base de dados, eleva-se a revocação para, digamos, 70–80%. Ainda sobrariam, neste caso hipotético, uns 20 a 30% de itens esquivos que o especialista em buscas provavelmente não conseguiria encontrar.

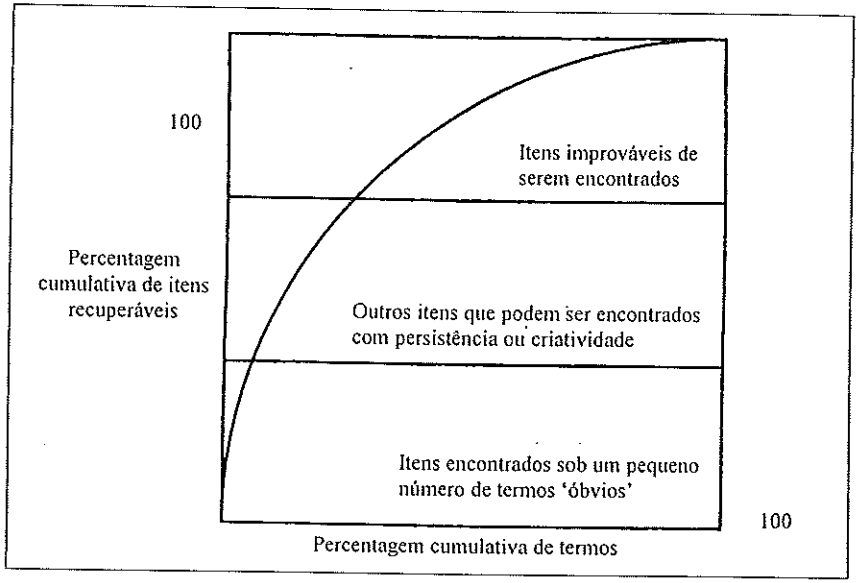


FIGURA 62
Dispersão de itens sob termos de indexação

Esta análise sobre simulações foi deliberadamente simplificada pelo fato de, em grande parte, ter suposto que uma busca teria apenas uma única faceta ou, pelo menos, seria uma busca feita num índice impresso onde só se pode consultar um termo de cada vez. A simulação de uma busca numa base de dados em linha, que comumente envolve mais de uma faceta, será um pouco mais complicada. Por exemplo, numa busca sobre pragas de insetos que atacam a soja, temos de admitir que só se recuperaria algum item se ele estivesse indexado sob um termo designativo de 'inseto' bem como sob um termo que indicasse 'soja'.

Albright (1979) realizou o estudo de recuperabilidade mais completo, utilizando, porém, uma única fonte, o *Index Medicus*. Martyn (1967) e Martyn e Slater (1964) examinaram a dispersão de material relevante sob termos de indexação em vários serviços impressos, e Bourne (1969a,b) também deu atenção à dispersão em seus estudos sobre a *Bibliography of Agriculture*. Carroll (1969) estudou a dispersão da literatura de virologia no *Biological Abstracts* e encontrou trabalhos sobre essa área dispersos em 20 seções dessa ferramenta além das que se referiam diretamente à virologia. O'Connor e Meadows (1968) encontraram dispersão semelhante da literatura de astronomia no *Physics Abstracts*.

Davison e Matthews (1969) examinaram a recuperabilidade de itens sobre computadores em espectrometria de massa em 11 serviços, bem como a cobertura desse assunto por parte desses serviços. Thorpe (1974) calculou a revocação e a precisão de buscas sobre reumatologia no *Index Medicus*, e Virgo (1970) utilizou o tema oftalmologia para comparar a recuperação da base de dados MEDLARS com a de seu principal produto, o *Index Medicus*. Jahoda e Stursa (1969) compararam as possibilidades de recuperação de um índice de assuntos de 'entrada única' com um índice baseado em palavras-chave dos títulos, Yerkey (1973) comparou as possibilidades de recuperação de um índice KWIC com o *Engineering Index* e o *Business Periodicals Index*, e Farradane e Yates-Mercer (1973) avaliaram o *Metals Abstracts Index* por meio de buscas simuladas.

Um método de avaliação dos índices impressos consiste em empregar sujeitos humanos no desempenho de tarefas de localização. Índices diferentes podem então ser comparados em termos de sucesso e eficiência (por exemplo, tempo de busca) no desempenho da tarefa. Um estudo desse tipo é relatado por van der Meij (2000), que compara diferentes formatos de apresentação de índices impressos do tipo incluído no final dos livros.

Olason (2000) também trata da usabilidade dos índices impressos, limitando seu estudo aos índices de livros. Seu estudo incluiu a cooperação de voluntários a quem foram atribuídas tarefas de localização de informações que exigiam o emprego de determinados índices. Foram registrados os tempos exigidos para completar as tarefas, bem como os caminhos de acesso usados pelos participantes; foram também solicitados a fazer comentários. Olason ocupa-se fundamentalmente dos efeitos do formato do índice na eficiência de uso.

Os estudos mais completos sobre desempenho da recuperação em índices impressos foram relatados por Keen (1976), tendo como assunto a biblioteconomia e a ciência da informação. As buscas foram feitas por estudantes e os resultados avaliados quanto a revocação, precisão e tempo de busca. Keen (1977b) também apresentou uma análise de estratégias de busca aplicadas a índices impressos.

Conaway (1974) desenvolveu um valor quantitativo único para expressar o mérito de um índice impresso, o Coeficiente de Usabilidade de Índices (CUI), o qual reflete quanto tempo leva um especialista em buscas para localizar as informações bibliográficas completas de determinado item. Uma busca temática era considerada bem-sucedida se o especialista conseguia encontrar um item que era de antemão conhecido como 'relevante' sobre um assunto dado. Se o item fosse localizado, registrava-se o tempo despendido para encontrar os dados bibliográficos completos. Empregando-se os métodos de Conaway, é possível atribuir escores numéricos a diferentes índices extraindo-se a média dos resultados obtidos sobre um número de assuntos por diversos especialistas em bus-

cas. O CUI é basicamente uma medida de custo–eficácia. No entanto, é uma medida muito medíocre, pois a eficácia é determinada exclusivamente com base na recuperação ou não-recuperação de um único item conhecido. Uma medida muito melhor de custo–eficácia é o custo unitário (em dinheiro ou em tempo do usuário) *por item relevante recuperado.*

Previsibilidade

A análise aqui feita sobre avaliação da recuperabilidade adotou um pressuposto importante: o de que é possível reconhecer um item 'relevante' a partir das informações sobre esse item contidas na base de dados. Estas informações compreendem:

1. O título do item
2. O título mais uma lista de termos de indexação
3. O título mais um resumo
4. O título mais os termos mais o resumo

Em geral, quanto mais extensa for a representação mais pistas fornecerá sobre se um item será ou não de interesse para o usuário. A informação mínima proporcionada por uma base de dados seria o título do item. O grau com que o título reflete satisfatoriamente o conteúdo temático depende em grande medida do tipo de publicação. Em geral, os artigos de periódicos científicos costumam trazer títulos bastante descritivos, enquanto, no outro extremo, as matérias de jornais apresentam títulos atraentes e que prendem a atenção, mas não são muito descritivos de seu conteúdo. As publicações técnicas ou comerciais também se inclinam pelo título atraente: o *Journal of Metals* apresenta títulos muito descritivos, sendo menos provável encontrá-los numa revista como *Iron Age.*

Os títulos, evidentemente, não são apresentados isoladamente. Num índice impresso, por exemplo, o título se situa no contexto do termo de indexação sob o qual aparece. O título 'Uma complicação rara da tuberculose' pouco nos diz a respeito do conteúdo de um artigo, mesmo que apareça sob o cabeçalho TUBERCULOSE PULMONAR. Se o mesmo título aparecesse sob o cabeçalho AMILOIDOSE ter-se-ia, no entanto, uma idéia muito melhor sobre seu conteúdo temático. Em alguns casos, também, o título do periódico (ou do livro) onde aparece um artigo pode dar uma pista de seu conteúdo temático. Assim, um artigo intitulado 'Efeitos sobre a apresentação de informações' faz pouco sentido visto isoladamente. Dentro de um livro intitulado *Editoração eletrônica*, por outro lado, esse título torna seu conteúdo muito mais previsível.

É algo raro um índice impresso incluir uma lista completa dos termos de indexação associados a um item (embora os índices da *Excerpta Medica* o façam), mas, comumente, é possível gerar uma lista dessas numa saída impressa de uma base de dados em linha, cuja indexação tenha sido feita por seres humanos. A combinação de título e termos de indexação é bastante eficaz para indicar de que trata uma publicação.

Os resumos, naturalmente, são os melhores indicadores de conteúdo. O principal critério para aferir sua qualidade é seu desempenho como preditores do conteúdo dos documentos.

Para testar a utilidade de várias formas de sucedâneos de documentos como indicadores do conteúdo destes, é preciso que se apresentem aos usuários de um sistema de recuperação (ou pessoas que estejam no lugar desses usuários em condições experimentais) várias representações de documentos de extensão crescente. Por exemplo, suponhamos que uma busca numa base de dados tenha recuperado 30 registros. As representações desses itens seriam apresentadas ao solicitante da busca numa seqüência de registros de extensão crescente, com os resultados mostrados no final deste parágrafo. Nessa situação hipotética, o solicitante, ao examinar o texto integral dos artigos de periódicos, considera 14 como relevantes e 16 como não relevantes. Suas previsões de relevância melhoraram à medida que crescia a extensão da representação do documento, embora o acréscimo de termos de indexação ao resumo não tenha sido diferente do uso do resumo sozinho. Até mesmo o melhor sucedâneo (título mais resumo) não foi perfeito: sub-representou os itens relevantes e super-representou os irrelevantes.

Registro apresentado	*Itens apresentados*	*Itens considerados nitidamente irrelevantes*	*Itens considerados relevantes ou possivelmente relevantes*
Título do artigo	30	12	18
Título do artigo mais título do periódico	30	13	17
Título do artigo (e do periódico) mais lista de termos de indexação	30	15	15
Título do artigo (e do periódico) mais resumo	30	18	12
Título do artigo (e do periódico) mais resumo e termos de indexação	30	18	12
Texto integral dos artigos	30	16	14

Pesquisas sobre o efeito de sucedâneos de documentos na previsibilidade da relevância foram feitas por diversos estudiosos, inclusive Rath et al. (1961a), Resnick (1961), Kent et al. (1967), Dym (1967), Shirey e Kurfeerst (1967), Saracevic (1969), Marcus et al. (1971) e Keen (1976). Marcus et al. demonstraram claramente que a 'indicatividade' de um sucedâneo de documento está diretamente relacionada à sua extensão em número de palavras. Por outro lado, é bem provável que haja uma extensão ideal que não seria econômico ultrapassar, pelo menos para fins de previsibilidade. Hagerty (1967), por exemplo, verificou que, embora a extensão de um resumo melhorasse as previsões de relevância, o efeito do aumento da extensão do resumo era surpreendentemente discreto.

Pesquisas sobre a utilidade dos resumos na previsão da relevância dos documentos pressupõem, geralmente, que o resumo é uma peça independente do documento, que aparece num serviço de resumos impressos ou na saída de uma operação de recuperação. Thompson (1973), no entanto, estudou a utilização e a utilidade dos resumos que acompanham os documentos (no início de artigos de periódicos ou de relatórios técnicos). Ele coletou dados sobre decisões quanto às atitudes tomadas por engenheiros e cientistas de três laboratórios militares em relação a documentos que passavam por suas mesas no curso das atividades rotineiras durante um período de quatro semanas. Não pôde confirmar se as decisões quanto às atitudes tomadas em relação aos documentos que continham resumos ocorriam mais rapidamente do que as decisões em relação àqueles sem resumos. Além disso, quando os sujeitos do experimento voltaram a receber cópias do documento, posteriormente, para um 'reexame ponderado', suas decisões anteriores quanto à relevância de documentos que continham resumos não apresentaram maior probabilidade de concordância com as decisões posteriores mais ponderadas do que fora constatado para os documentos sem resumos. Estes resultados não lançam dúvida sobre a utilidade dos resumos de per si ou mesmo sobre a utilidade dos resumos que acompanham artigos ou relatórios (uma vez que são freqüentemente adotados ou modificados por serviços secundários), mas realmente sugerem que os resumos podem ter um uso limitado nas decisões de seleção inicial. Muitos dos indivíduos a quem são enviadas publicações preferem julgar a relevância delas para seus interesses atuais passando os olhos no texto, examinando as tabelas ou figuras, ou mesmo checando as referências bibliográficas (por exemplo, para conferir se foram citados!).

A qualidade dos resumos como tais é examinada no capítulo 8, enquanto o tópico relativo à elaboração automática de resumos é tratado no capítulo 15.

Em estudo mais recente, Salton et al. (1997) compara resumos de documentos complexos elaborados automaticamente com base na extração em parágrafos com processo similar, feito por seres humanos, de extração em parágrafos. Eles justificam o método automático com o argumento de que um resumo elaborado automaticamente é tão provável que coincida com um resumo preparado por seres humanos quanto um resumo preparado por uma pessoa coincida com um resumo preparado por outra pessoa.

Processos semelhantes foram empregados na avaliação de traduções feitas automaticamente ou por seres humanos. Brew e Thompson (1994), por exemplo, argumentam que "as boas traduções tenderão a ser mais similares entre si do que as traduções ruins".

Atualidade

A atualidade ou 'presteza' é uma medida da velocidade com que novas publicações são incluídas num serviço de indexação/resumos. Trata-se de um critério que os usuários percebem imediatamente, pois a data de publicação de um índice impresso é conhecida e a data (ou pelo menos o ano) da primeira edição de cada item incluído consta de sua referência bibliográfica. A atualidade é menos aparente para usuários de sistemas em linha, mas ainda assim é perceptível.

Essa visibilidade é desastrosa porque comumente leva a conclusões que não são válidas. Uma tendência humana é perceber casos excepcionais, e outra tendência é dar às expectativas um peso indevido ao fazer um julgamento. O usuário de um volume de resumos impressos tem a oportunidade de examinar inúmeras referências ao mesmo tempo. Ao perceber que algumas correspondem a materiais publicados há talvez dois ou três anos, ele injustificadamente conclui que o serviço é em geral muito lento na identificação e processamento de novos itens.

Existem muitos motivos pelos quais a inclusão de uma referência num arquivo sofre atrasos. O intervalo entre a impressão de um periódico e seu recebimento pelo serviço secundário pode ser longo devido a razões de ordem geográfica ou econômica; por exemplo, um serviço norte-americano recebe os periódicos dos EUA pelo correio poucos dias depois de sua impressão, porém os periódicos estrangeiros podem levar de seis a sete semanas para chegar a esse país. Certos tipos de materiais, como os anais de eventos, são difíceis de localizar e, uma vez localizados, são de aquisição trabalhosa. Documentos escritos em certas línguas demandam mais do que o tempo médio para serem processados, devido à escassez de tradutores qualificados. Materiais 'periféricos', o que comumente significa materiais aparecidos em periódicos e outras publicações que não sejam examinados rotineiramente pelo serviço, tomam mais tempo para serem localizados do que os materiais da lista básica, uma vez que freqüentemente só são identificados quando da consulta a outros serviços secundários e, por isso, sofrem uma dupla série de atrasos no processamento. Alguns serviços contam com sistemas de processamento mais ágeis do que outros, e alguns atrasos são imputáveis à ineficiência do sistema. Quando um serviço de indexação/resumos é utilizado para notificação corrente, a avaliação que dele faz o usuário é influenciada pelo número de itens dos quais ele já tem conhecimento prévio e que constam do fascículo mais recente. A presença de alguns itens já conhecidos costuma estimular a confiança na eficiência do serviço, mas a presença de um número excessivo desses itens abala a confiança em sua atualidade.

Do ponto de vista do avaliador, a atualidade é muito atraente como critério de eficácia. A atualidade é relativamente fácil de medir, sendo incontestável quando medida, porque não depende de juízos subjetivos. A única influência que o avaliador exerce sobre a medida é na escolha das datas que serão usadas. No caso de bases de dados impressas, a data de aparecimento da referência é normalmente tida como a data de publicação do serviço. No caso de um serviço em linha, a data será aquela em que a referência foi incluída na base de dados, mas esta data nem sempre pode ser confirmada retrospectivamente. Uma solução possível é saber junto à editora qual o intervalo entre a data de distribuição da atualização da base de dados eletrônica e a de seu equivalente em versão impressa, e ajustar as medidas de conformidade com isso. Para a data de aparecimento

da publicação primária a que se refere o serviço secundário, o avaliador pode escolher entre a data efetiva de edição e a data em que a publicação se torna disponível.

A data efetiva da edição apresenta alguns problemas, pois raramente ela é fornecida com exatidão nos serviços secundários. Recorre-se a um exemplar da publicação e, na maioria dos casos, a data consignada refere-se somente ao mês mais próximo. O dia efetivo em que a publicação foi editada somente será conhecido se se fizer uma consulta à editora. Infelizmente, a data na capa de um periódico nem sempre é confiável, pois alguns são publicados no mês anterior ao mês nominal de edição, e a maioria aparece posteriormente à data nominal de edição.

A data em que a publicação se torna disponível é, em alguns países, uma alternativa que não apresenta problemas. Embora essa data — data de disponibilidade — realmente não meça a atualidade de um serviço, ela nos dá uma medida da atualidade efetiva do ponto de vista dos usuários do país onde são registradas as datas de disponibilidade. Nos EUA, as datas de disponibilidade de periódicos seriam a data de seu recebimento pela Library of Congress ou outra das bibliotecas nacionais. Essas datas aparecem carimbadas nas capas dos periódicos, ou provavelmente existam num registro de controle mantido pela biblioteca. Pode-se assim medir o intervalo de tempo entre a disponibilidade de um periódico nos Estados Unidos e a notificação de sua existência por algum serviço secundário. Estritamente falando, deve-se considerar a data de disponibilidade do serviço secundário ao invés da data de sua edição, porém raramente se verifica uma grande diferença entre elas. A atualização da base de dados eletrônica normalmente se dá antes da atualização do índice impresso ao qual se relaciona.

A coleta de dados implica a obtenção de uma amostra aleatória de itens extraídos do fascículo mais recente de um serviço secundário, anotando-se a data de edição (ou atualização) do serviço, a isto acrescentando os dados originais de edição ou disponibilidade, normalmente retirados de uma fonte diferente. Se se exigir, como convém freqüentemente, que seja feita uma análise por línguas, países de origem e formas de publicação (por exemplo, artigos de periódicos, teses e monografias), as dimensões da amostra deverão ser maiores do que seriam se se desejasse apenas uma estimativa global da atualidade.

A atualidade é provavelmente a característica de um serviço secundário mais fácil de ser medida. Também é provável que seja a menos importante. As editoras talvez se interessem pela atualidade como medida da eficácia de suas atividades, mas os usuários, embora freqüentemente expressem a vontade de que o serviço seja ágil, talvez se impressionem menos com isso. Quando se leva em consideração o intervalo de tempo decorrido entre a conclusão de uma pesquisa e sua primeira publicação, a demora adicional acarretada pela utilização de um serviço secundário para localizar essa pesquisa é relativamente pequena.

Normas

Em teoria, um método óbvio para fazer a avaliação de índices e resumos é compará-los com as normas existentes. No mundo anglófono, as normas pertinentes são:

ANSI/NISO Z39.14–1997 *Guidelines for abstracts* (reeditada em 2002)
ANSI/Z39.4–1984 *Basic criteria for indexes*
ISO 999 : 1996 *Guidelines for the content, organization, and presentation of indexes*
ISO 5963–1985 (E) *Methods for examining documents, determining their subjects, and selecting indexing terms*
BS 3700 : 1988 *Preparing indexes to books, periodicals, and other documents*
BS 6529 : 1984 *Examining documents, determining their subjects, and selecting indexing terms*

Observe-se que essas normas tendem a enfocar o produto ao invés do processo: os índices e os resumos ao contrário da indexação e da redação de resumos. Apenas a ISO 5963 e a BS 6529 focalizam o processo. Em virtude de estar voltadas para o aspecto mais difícil da indexação — decidir qual é realmente a 'matéria indexável' de um documento — são, por vários motivos, as mais úteis de todas as normas que lidam com a análise conceitual dos documentos.

Na realidade, embora essas diferentes publicações sejam editadas por organizações de normalização, é difícil considerá-las como verdadeiras normas. Uma norma de verdade deve ser exata (por exemplo, a norma relativa à composição de determinada liga) e de cumprimento obrigatório (por exemplo, a norma que especifica que determinado tipo de aço deve ter uma resistência à tração igual a x). É claro que a indexação e a redação de resumos são atividades que não são nem exatas nem de cumprimento obrigatório (exceto sob condições muito limitadas, como, por exemplo, os requisitos que sejam exigidos pelos editores de um periódico). A imprecisão e a evidente subjetividade da indexação estão bem-demonstradas no fato de a comissão de desenvolvimento de normas da NISO, incumbida de rever a ANSI Z39.4, não ter chegado a acordo quanto a uma norma e só ter conseguido produzir um relatório que "servissse como um recurso atual sobre indexação" (*Guidelines for indexes*, 1997). Por conseguinte, esse relatório simplesmente possui um número de relatório e não uma designação oficial de norma Z39. É difícil compreender tanta sutileza, pois, como se disse antes, normas de verdade não podem (e provavelmente não devem) ser impostas a atividades intelectuais, e a maioria das 'normas' tem outra denominação ('diretrizes' ou 'critérios'). Qualquer que seja a forma como sejam chamadas, essas normas não são suficientemente precisas para serem usadas na avaliação de índices ou resumos, ou na indexação e redação de resumos, exceto no nível mais superficial. Ademais, as normas de indexação se concentram basicamente nos índices impressos em geral e nos índices de final de livro, em particular.

Outros aspectos concernentes à avaliação

Vários outros métodos foram empregados para analisar o desempenho e o

uso de índices impressos. Por exemplo, Torr et al. (1966) descrevem quatro métodos que podem ser adotados para 'observar' os usuários de índices: 1) fazer com que o usuário mantenha um registro escrito dos processos de raciocínio e da estratégia que emprega ao fazer uma busca, 2) fazer com que o especialista em buscas utilize um gravador de fita com a mesma finalidade,* 3) fazer com que um observador acompanhe a busca, e 4) empregar a observação humana combinada com uma câmara para estudar como os índices são utilizados. Esses pesquisadores verificaram ser difícil conseguir a cooperação dos 'usuários reais' com esses estudos, o que também foi a experiência de Hall (1972).

Outros pesquisadores empregaram entrevistas ou questionários para obter as opiniões de usuários relativas a vários serviços de indexação/resumos, inclusive Hall (1972a,b), Keen (1976), Drage (1969) e Cluley (1968).

Ao tratar da recuperabilidade, este capítulo deteve-se mais nos serviços impressos de indexação e resumos do que na recuperação em bases de dados eletrônicas. Isso reflete parcialmente o foco do presente livro: indexação e redação de resumos ao invés de outros aspectos da recuperação da informação. Evidentemente, os métodos adotados para estudar cobertura, previsibilidade e atualidade são pertinentes a todos os tipos de bases de dados, impressas ou eletrônicas. Os estudos de cobertura e de atualidade são completamente objetivos, e os estudos de previsibilidade um tanto menos. Os estudos de recuperabilidade são inerentemente subjetivos, pois dependem de decisões humanas a respeito de quais itens são relevantes (ou pertinentes)** e quais não são. Ao estudar a eficácia da recuperação, precisa-se utilizar uma medida que reflita a proporção dos itens relevantes que são recuperados durante uma busca (coeficiente de revocação), bem como alguma medida do *custo* da recuperação dessa parcela da literatura relevante. O coeficiente de precisão é comumente empregado como uma medida indireta do custo, pois reflete o número de itens que o usuário de algum modo deve examinar a fim de identificar *n* itens que lhe sejam úteis. Outra medida indireta do custo é a extensão esperada da busca, descrita por Cooper (1968). Naturalmente, pode-se medir o custo de uma maneira mais direta levando-se em conta todos os custos da busca, inclusive o tempo do especialista em buscas e os custos de acesso à base de dados (ver, por exemplo, Elchesen, 1978). O custo da busca será então relacionado ao número de itens relevantes (ou pertinentes, ou úteis ou 'novos') recuperados; o 'custo por referência relevante recuperada' é uma boa medida da relação custo–eficácia da busca.

As medidas de eficácia, como a revocação e a precisão (ou outras descritas, por exemplo, por Robertson, 1969), são aplicáveis a estudos de recuperação em qualquer tipo de base de dados, tanto em formato impresso quanto eletrônico.

---

* Keen (1977b) também utilizou esta técnica.

** A questão da relevância/pertinência foi examinada por muitos autores. Ver, por exemplo, Wilson (1973), Swanson (1986), Lancaster e Warner (1993), e Mizzaro (1998).

No entanto, quando estudamos a eficácia da recuperação, torna-se bastante difícil isolar os efeitos da indexação/redação de resumos de outros fatores, tais como o vocabulário da base de dados, as estratégias de busca empregadas e a interação usuário/sistema. Isso foi mencionado de passagem no capítulo 6. Foge aos propósitos deste livro descrever detalhadamente a metodologia da avaliação (mensuração da precisão, cálculo da revocação, análise de diagnóstico para determinar as causas exatas das falhas de revocação e precisão). Este assunto é tratado de modo completo em Lancaster e Warner (1993).

As avaliações de serviços impressos de indexação/resumos, ou seus equivalentes eletrônicos, são menos comuns atualmente do que antes, em parte porque agora se dedica mais atenção aos estudos relacionados com a Rede (por exemplo, avaliações de mecanismos de buscas ou sítios da Rede). Não obstante, ainda se publicam algumas avaliações. Exemplos disso são os trabalhos de Brown et al. (1999), que compararam a cobertura do *Current Index to Journals in Education* com o *Education Index*, e os de Brettle (2001), que comparou diferentes bases de dados do ponto de vista da cobertura de informações sobre a reabilitação de portadores de doença mental grave, e Green (2001), que incluiu a cobertura (junto com a atualidade) numa avaliação de bases de dados de periódicos de música. Ambos concluíram que seriam necessárias múltiplas bases de dados para uma cobertura adequada desses assuntos.

Azgaldov (1969) identificou alguns critérios que podem ser empregados para avaliar a qualidade de índices impressos. Tais critérios são: *adequação* (que abrange toda uma gama de propriedades, que incluem cobertura, características do vocabulário usado na indexação, bem como certos fatores dependentes da indexação, como a exaustividade e a coerência), *generalidade* (que diz respeito essencialmente à diversidade de buscas que podem ser feitas), *ergonomicidade* (facilidade de uso), *presteza* (quão atualizada é a fonte), e *custo*. Ele ressalta, muito corretamente, que:

> O mais eficiente índice impresso será um fracasso para os usuários, se seu parâmetro de conveniência [ergonomicidade e presteza] for baixo, e, vice-versa, um índice que for simples e fácil de usar granjeará ampla popularidade mesmo que seu desempenho na recuperação não seja muito alto (p. 281).

Esta citação serve como um bom intróito ao capítulo 10, que trata das características de vários serviços impressos de indexação e resumos.

Capítulo 10

Métodos adotados em serviços impressos de indexação e resumos

A finalidade deste capítulo é expor vários métodos de implementação de serviços de indexação e resumos em formato impresso. Em particular, considera as bases de dados impressas em função de suas propriedades como ferramentas de recuperação da informação.*

Identificam-se, basicamente, dois métodos principais de organização dessas ferramentas. Num deles, as entradas aparecem sob cabeçalhos de assuntos ou descritores relativamente específicos dispostos em ordem alfabética. As entradas podem repetir-se sob mais de um cabeçalho e/ou são empregadas remissivas para ligar cabeçalhos relacionados entre si. Neste método não há necessidade de índice de assuntos, porém, serão necessários outros tipos de índices, principalmente de autores.

No outro método, utiliza-se uma forma de classificação: as entradas são dispostas sob números de classificação altamente específicos ou agrupadas sob categorias temáticas relativamente genéricas (possivelmente com subcategorias). Em qualquer um dos casos, há necessidade de índices de assuntos que permitam abordagens alternativas ou acesso mais específico ao conteúdo temático.

Índices alfabético-específicos

Um dos melhores exemplos deste método é a edição mensal do *Index Medicus* e sua edição acumulada, o *Cumulated Index Medicus* (figura 63). Várias características deste índice merecem atenção:

1. Note-se como são usados subcabeçalhos que oferecem maior especificidade.
2. Como não são incluídos resumos, torna-se viável repetir a referência bibliográfica sob mais de um cabeçalho. Por exemplo, a primeira entrada sob o subcabeçalho *administration & dosage* (figura 63) provavelmente será duplicada sob OSTEOARTHRITIS.
3. A combinação de cabeçalho, subcabeçalho e título do artigo normalmente proporciona uma imagem relativamente clara daquilo de que trata um item.

* Na realidade, estas ferramentas impressas são muito menos utilizadas hoje em dia do que o eram quando foram publicadas as edições anteriores deste livro. Muitas bibliotecas cancelaram as assinaturas das versões impressas, dando preferência ao acesso em linha e, em alguns casos, as edições impressas, ou partes delas, foram interrompidas pelas respectivas editoras.

4. Dois tipos de remissivas aparecem no índice impresso: *see* [ver] é usada para ligar termos considerados sinônimos ou quase sinônimos e *see related* [ver em relação a isto] para ligar termos intimamente relacionados. Para se obter, contudo, um quadro completo da rede de associações entre os termos utilizados, é preciso consultar dois outros instrumentos: *Medical subject headings* (*MeSH*) e *MeSH tree structures*. A figura 64 mostra um exemplo de uma página do *MeSH*. Observe-se como o *MeSH* apresenta a remissiva *see* (e sua recíproca *X* bem como as remissivas *see related* (recíproca *XR*) empregadas para ligar dois termos semanticamente relacionados, normalmente pertencentes a hierarquias diferentes. Talvez ainda mais importante, a cada cabeçalho do *MeSH* é atribuído um ou mais números de classificação para indicar onde ele aparece nas estruturas hierárquicas em árvore (figura 65). Assim, embora o vocabulário utilizado pela National Library of Medicine seja bastante rico em associações, o *Index Medicus* não é auto-suficiente, pois nele não aparecem as associações. Portanto, é uma fonte útil em buscas relativamente específicas, mas difícil de usar em buscas de caráter mais genérico que exijam a consulta a muitos cabeçalhos diferentes.

A figura 66 mostra exemplos de entradas do índice de autores do *Cumulated Index Medicus*. Observe-se que se tem aqui uma estrutura totalmente auto-suficiente, pois não se trata de um índice da seção de assuntos. Na realidade, para qualquer item encontrado no índice de autores é freqüentemente bastante difícil determinar quais são os cabeçalhos de assuntos sob os quais ele aparece. Note-se também que o índice de autores, ao contrário da seção de assuntos, arrola todos os autores de cada artigo e traz o título do artigo na língua original (pelo menos para línguas escritas com alfabeto romano), não em tradução. O *Cumulated Index Medicus* não é mais publicado, mas o *Index Medicus* mensal, sim.

Os vários índices impressos editados pela H.W. Wilson Co. (dos quais são bons exemplos o *Reader's Guide to Periodical Literature* e o *Library Literature*) são em muitos aspectos similares ao *Index Medicus*, pois utilizam cabeçalhos específicos com subcabeçalhos e incorporam remissivas do tipo *see*. Diferem do *Index Medicus* por adotarem muito mais remissivas *see also* [ver também] para ligar termos semanticamente relacionados, tornando um tanto mais fácil a realização de buscas genéricas que envolvam vários cabeçalhos diferentes. Por exemplo (ver figura 67), o usuário que consulte o termo MAGNETOHYDRODYNAMICS (no *Applied Science and Technology Index*) é informado de que deve procurar também sob PLASMA, PLASMA WAVES e SYNCHROTRON RADIATION.

O *Engineering Index* também organizava suas entradas sob cabeçalhos específicos e subcabeçalhos e incluía tanto remissivas do tipo *see* quanto do tipo *see also*. Hoje, porém, as referências são arranjadas sob descritores sem subcabeçalhos (ver figura 68).

A principal diferença entre este índice e os que foram anteriormente exemplificados é, obviamente, o fato de incluir resumos. Cada resumo recebe um nú-

mero de identificação exclusivo. O índice de autores, então, é um verdadeiro índice do arranjo por assuntos, remetendo do nome do autor para os números dos resumos aos quais esse nome está associado. Ademais, como muitas entradas acumular-se-ão sob os cabeçalhos do volume anual, também existe um índi-

CYCLONAMINE see ETHAMSYLATE

CYCLOOXYGENASE see PROSTAGLANDIN-ENDOPEROXIDE SYNTHASE

CYCLOOXYGENASE INHIBITORS

see related
ANTI-INFLAMMATORY AGENTS, NON-STEROIDAL

ADMINISTRATION & DOSAGE

Meloxicam in osteoarthritis: a 6-month, double-blind comparison with diclofenac sodium. Hosie J, et al. Br J Rheumatol 1996 Apr;35 Suppl 1:39–43

Peri-operative administration of rectal diclofenac sodium. The effect on renal function in patients undergoing minor orthopaedic surgery. Irwin MG, et al. Eur J Anaesthesiol 1995 Jul;12(4):403–6

Transdermal modification of platelet function: an aspirin patch system results in marked suppression of platelet cyclooxygenase. McAdam B, et al. J Pharmacol Exp Ther 1996 May;277(2):559–64

[What dose of aspirin should be prescribed in patients with coronary disease?] Montalescot G, et al. Ann Cardiol Angeiol (Paris) 1995 Oct;44(8):469–72 (Eng. Abstr.) (Fre)

ADVERSE EFFECTS

Tolerability of imidazole salycilate in aspirin-sensitive patients. Senna OE, et al. Allergy Proc 1995 Sep-Oct; 16(5):251–4

Doppler echocardiographic findings of indomethacin-induced occlusion of the fetal ductus arteriosus. Takahashi Y, et al. Am J Perinatol 1996 Jan; 13(1):15–8

Impact of preexisting health conditions on the outcome of an adverse drug reaction alerting program: gastrointestinal disorders before piroxicam and sulindac therapy. Rawson NS. Ann Pharmacother 1995 Jul-Aug;29(7–8):676–80

Abnormally high platelet activity after discontinuation of acetylsalicylic acid treatment. Beving H, et al. Blood Coagul Fibrinolysis 1996 Jan;7(1):80–4

Effects of free radical scavengers on indomethacin-induced aggravation of gastric ulcer in rats. Naito Y, et al. Dig Dis Sci 1995 Sep;40(9):2019–21

Effect of ketorolac tromethamine on bleeding and on requirements for analgesia after total knee arthroplasty [letter; comment] Dodenhoff RM. J Bone Joint Surg Am 1996 Jun;78(6):968. Comment on: J Bone Joint Surg Am 1995 Jul;77(7):998–1002.

Effect of ketorolac tromethamine on bleeding and on requirements for analgesia after total knee arthroplasty [letter; comment] Linville D. J Bone Joint Surg Am 1996 Jun;78(6):967–8. Comment on: J Bone Joint Surg Am 1995 Jul;77(7):998–1002.

NSAIDs, Cox-2 inhibitors, and the gut [letter; comment] [see comments] Bennett A, et al. Lancet 1995 Oct 21; 346(8982):1105. Comment on: Lancet 1995 Aug 26;346(8974):521–2. Comment in: Lancet 1995 Dec 16;346(8990):1629.

NSAIDs, Cox-2 inhibitors, and the gut [letter; comment] Vane JR. Lancet 1995 Oct 21;346(8982):1105–6. Comment on: Lancet 1995 Aug 26;346(8974):521–2.

[Sensitivity to acetylsalicylic acid] Elverland HH. Tidsskr Nor Laegeforen 1996 Feb 28;116(6):754–6 (22 ref.) (Eng. Abstr.) (Nor)

CHEMICAL SYNTHESIS

Diarylspiro[2.4]heptenes as orally active, highly selective cyclooxygenase-2 inhibitors: synthesis and structure-activity relationships. Huang HC, et al. J Med Chem 1996 Jan 5;39(1):253–66

Design, synthesis, and biochemical evaluation of N-substituted maleimides as inhibitors of prostaglandin endoperoxide synthases. Kalgutkar AS, et al. J Med Chem 1996 Apr 12;39(8):1692–703

CHEMISTRY

Hydroxylamine analogs of 2,6-di-t-butylphenols: dual inhibitors of cyclooxygenase and 5-lipoxygenase or selective 5-lipoxygenase inhibitors. Kramer JB, et al. Bioorg Med Chem 1995 Apr;3(4):403–10

Involvement of arginine 120, glutamate 524, and tyrosine 355 in the binding of arachidonate and 2-phenylpropionic acid inhibitors to the cyclooxygenase active site of ovine prostaglandin endoperoxide H synthase-1. Bhattacharyya DK, et al. J Biol Chem 1996 Jan 26;271(4):2179–84

A single amino acid difference between cyclooxygenase-1 (COX-1) and -2 (COX-2) reverses the selectivity of COX-2 specific inhibitors. Gierse JK, et al. J Biol Chem 1996 Jun 28;271(26):15810–4

The structural basis of aspirin activity inferred from the crystal structure of inactivated prostaglandin H2 synthase [see comments] Loll PJ, et al. Nat Struct Biol 1995 Aug; 2(8):637–43. Comment in: Nat Struct Biol 1995 Aug;2(8):605–6.

IMMUNOLOGY

Oral aspirin and ibuprofen increase cytokine-induced synthesis of IL-1 beta and of tumour necrosis factor-alpha ex vivo. Endres S, et al. Immunology 1996 Feb;87(2):264–70

METABOLISM

Indomethacin, esculetin and nordihydroguaiaretic acid modify arachidonate biosynthesis in rat adrenocortical cells. de Gómez Dumm NT, et al. Acta Physiol Pharmacol Ther Latinoam 1995;45(3):155–64

Tyrosine kinase inhibitors prevent cytokine-induced expression of iNOS and COX-2 by human islets. Corbett JA, et al. Am J Physiol 1996 Jun;270(6 Pt 1):C1581–7

Synthesis and use of iodinated nonsteroidal antiinflammatory drug analogs as crystallographic probes of the prostaglandin H2 synthase cyclooxygenase active site. Loll PJ, et al. Biochemistry 1996 Jun 11;35(23):7330–40

In vivo inhibition profile of cytochrome P450TB (CYP2C9) by (+/-)-fluvastatin. Transon C, et al. Clin Pharmacol Ther 1995 Oct;58(4):412–7

Reactions of prostaglandin endoperoxide synthase and its compound I with hydroperoxides. Baković M, et al. J Biol Chem 1996 Jan 26;271(4):2048–56

Effects of indomethacin and arachidonic acid on sister chromatid exchange induction by styrene and styrene-7,8-oxide. Lee SH, et al. Mutat Res 1995 Oct; 348(2):93–9

PHARMACOKINETICS

The pharmacokinetic and pharmacodynamic interactions between the 5-lipoxygenase inhibitor zileuton and the cyclo-oxygenase inhibitor naproxen in human volunteers. Awni WM, et al. Clin Pharmacokinet 1995;29 Suppl 2:112–24

Pharmacokinetics and pharmacodynamics of tepoxalin after single oral dose administration to healthy volunteers. Waldman SA, et al. J Clin Pharmacol 1996 May;36(5):462–8

PHARMACOLOGY

Inhaled lysine acetylsalicylate (L-ASA) attenuates histamine-induced bronchoconstriction in asthma. Crimi N, et al. Allergy 1996 Mar;51(3):157–63

Influence of indomethacin on bone turnover related to orthodontic tooth movement in miniature pigs. Giunta D, et al. Am J Orthod Dentofacial Orthop 1995 Oct; 108(4):361–6

High-dose tumor necrosis factor alpha produces an impairment of hamster diaphragm contractility. Attenuation with a prostaglandin inhibitor. Wilcox P, et al. Am J Respir Crit Care Med 1996 May;153(5):1611–5

Effect of flunixin meglumine on endogenous prostaglandin F2 alpha secretion during cloprostenol-induced abortion in mares. Daels PF, et al. Am J Vet Res 1995 Dec; 56(12):1603–10

FIGURA 63
Exemplo de entradas do *Cumulated Index Medicus* (1996)

Receptors, Cyclic AMP

D12.776.543.750.720.700.150 D12.776.543.750.810.150
77
see related
Cyclic AMP-Dependent Protein Kinases
Cyclic AMP Receptor Protein
X cAMP Receptors
X Cyclic AMP Receptors
XR Cyclic AMP

Receptors, Cytoadhesin

D12.776.543.750.705.408.460+ D24.611.834.408.460+
90
X Receptors, Extracellular Matrix Glycoprotein
XR Extracellular Matrix Proteins

Receptors, Cytokine

D12.776.543.750.705.852+ D24.611.834.852+
94
X Cytokine-Receptors

Receptors, Cytoplasmic and Nuclear

D12.776.826+
94
see related
Transcription Factors
X Cytoplasmic and Nuclear Receptors
X Cytosolic and Nuclear Receptors
X Nuclear and Cytoplasmic Receptors

Receptors, delta see Receptors, Opioid, delta

Receptors, delta Opioid see Receptors, Opioid, delta

Receptors, Diazepam see Receptors, GABA-A

Receptors, Diiodotyrosine see Receptors, Thyroid Hormone

Receptors, Dioxin see Receptors, Aryl Hydrocarbon

Receptors, Dopamine

D12.776.543.750.600.300.400+ D12.776.543.750.720.300.300.400+
77
X Dopamine Receptors
XR Dopamine
Receptors, Dopamine/agonists see Dopamine Agonists
Receptors, Dopamine/antagonists & inhibitors see Dopamine Antagonists

Receptors, Dopamine D1

D12.776.543.750.600.300.400.400 D12.776.543.750.720.300.300.400.400
93; DOPAMINE-D1 RECEPTOR was indexed under RECEPTORS, DOPAMINE 1982-92
X Dopamine D1 Receptors

Receptors, Dopamine D2

D12.776.543.750.600.300.400.500 D12.776.543.750.720.300.300.400.500
93; DOPAMINE-D2 RECEPTOR was indexed under RECEPTORS, DOPAMINE 1982-92
X Dopamine D2 Receptors

FIGURA 64
Exemplo de entradas do *Medical subject headings* (1996)

Leukemia			
Leukemia by Immunologic Marker (Non MeSH)			
Leukemia, B-Cell			
Leukemia, B-Cell, Chronic			
Leukemia, B-Cell, Chronic	C4.557.337.150.125.250	C4.557.337.	
Leukemia, Pre-B-Cell	C4.557.337.150.125.650		
Leukemia, Mixed-Cell	C4.557.337.150.500	C4.557.337.	
Leukemia, Null-Cell	C4.557.337.150.550	C4.557.337.	
Leukemia, T-Cell	C4.557.337.150.800		
Leukemia, T-Cell, Acute	C4.557.337.150.800.100	C4.557.337.	
Leukemia-Lymphoma, T-Cell, Acute, HTLV-I-Associated	C4.557.337.150.800.100.300	C2.782.815, C20.673.483.	C4.557.337.
Leukemia, T-Cell, Chronic	C4.557.337.150.800.250	C4.557.337.	
Leukemia, T-Cell, HTLV-II-Associated	C4.557.337.150.800.350	C2.782.815, C20.673.483.	C4.557.337.
Leukemia, Experimental	C4.557.337.372	C4.619.531	
Avian Leukosis	C4.557.337.372.216	C2.782.815, C4.619.531, C4.935.320	C3.928.120 C4.619.935, C22.131.94
Leukemia L1210	C4.557.337.372.594	C4.619.531.	
Leukemia L5178	C4.557.337.372.602	C4.619.531.	
Leukemia P388	C4.557.337.372.782	C4.619.531.	
Leukemia, Feline	C4.557.337.385	C2.782.815.	C22.180.300
Leukemia, Hairy Cell	C4.557.337.415	C15.604.515.	C20.683.515.
Leukemia, T-Cell, HTLV-II-Associated	C4.557.337.415.700	C2.782.815, C20.673.483.	C4.557.337.
Leukemia, Lymphocytic	C4.557.337.428	C15.604.515.	C20.683.515.
Leukemia, Lymphocytic, Acute	C4.557.337.428.511		
Leukemia, B-Cell, Acute	C4.557.337.428.511.100	C4.557.337.	
Leukemia, CALLA-Positive	C4.557.337.428.511.225		
Leukemia, Lymphocytic, Acute, L1	C4.557.337.428.511.400		
Leukemia, Lymphocytic, Acute, L2	C4.557.337.428.511.410		
Leukemia, Mixed-Cell	C4.557.337.428.511.500	C4.557.337.	
Leukemia, Null-Cell	C4.557.337.428.511.550	C4.557.337.	
Leukemia, T-Cell, Acute	C4.557.337.428.511.800	C4.557.337.	
Leukemia-Lymphoma, T-Cell, Acute, HTLV-I-Associated	C4.557.337.428.511.800.300	C2.782.815, C20.673.483.	C4.557.337.
Leukemia, Lymphocytic, Chronic	C4.557.337.428.550		
Leukemia, B-Cell, Chronic	C4.557.337.428.550.250	C4.557.337.	
Leukemia, Prolymphocytic	C4.557.337.428.550.675		
Leukemia, T-Cell, Chronic	C4.557.337.428.550.800	C4.557.337.	
Leukemia, Mast-Cell	C4.557.337.440	C4.557.337.	
Leukemia, Myeloid	C4.557.337.539	C15.378.190. C20.683.515.	C15.604.515.

FIGURA 65
Exemplo de entradas da estrutura hierárquica (*Tree structures*) do *Medical subject headings* (1996)

ce mais específico de assuntos. A figura 69 mostra entradas do índice de assuntos de 1993, que emprega tanto descritores controlados (em tipo negrito) quanto termos de texto livre (em tipo normal). O índice refere-se às entradas tanto no volume anual (números que começam com A) quanto nos fascículos mensais (números que começam com M). Observe-se como uma das entradas da figura 69 relaciona-se com a entrada 073654 da figura 68, proporcionando acesso a este item sob o ponto de acesso alternativo BEAM PLASMA INTERACTIONS.

Muitos dos índices impressos (mas de modo algum todos) baseiam-se em alguma forma de vocabulário controlado — um tesauro ou uma lista de cabeçalhos de assuntos. O vocabulário utilizado pelo *Engineering Index* é o *Engineering Index thesaurus*. Tais vocabulários controlados são de grande valia para quem estiver consultando o índice impresso, principalmente em casos onde o próprio índice inclui pouca estrutura de remissivas, como acontece no *Index Medicus*.

Colquhoun J. Dental caries among children in New Zealand [letter; comment] Community Dent Oral Epidemiol 1995 Dec;23(6):381. Comment on: Community Dent Oral Epidemiol 1994 Aug;22(4):226-30.
Colquhoun JP. Heartsink revisited. Aust Fam Physician 1995 Oct;24(10):1964-5
Colquhoun JP. That was the week that was. Aust Fam Physician 1996 Aug;25(8):1333-4
Colquhoun JP. The index theory and the magic of medicine. Aust Fam Physician 1996 Jun;25(6):978-9 .
Colquhoun K see **Mahmood R**
Colquhoun KO, Timms S, Fricker CR. Detection of Escherichia coli in potable water using direct impedance technology. J Appl Bacteriol 1995 Dec;79(6):635-9
Colquhoun MC, Waine C, Monaghan MJ, Struthers AD, Mills PG. Investigation in general practice of patients with suspected heart failure. How should the essential echocardiographic service be delivered? [editorial] [see comments] Br Heart J 1995 Oct;74(4):335-6. Comment in: Heart 1996 Jun;75(6):642; discussion 643. Comment in: Heart 1996 Jun;75(6):642-3. Comment in: Heart 1996 Jun;75(6):643. Comment in: Heart 1996 Jun;75(6):643-4.
Colquhoun MC, Waine C, Monaghan MJ, Struthers AD, Mills PG. Investigation in general practice of patients with suspected heart failure: how should the essential echocardiographic service be delivered? [editorial] Br J Gen Pract 1995 Oct;45(399):517-9
Colquhoun S see **Swanson C**
Colquhoun SD. Hepatitis C. A clinical update. Arch Surg 1996 Jan;131(1):18-23 (49 ref.)
Colquhoun SD see **Imagawa DK**
Colquhoun-Flannery W, Carruth JA. Diet-modified sex hormone metabolism: is this the way forward in recurrent respiratory papillomatosis and squamous carcinoma prophylaxis? J Laryngol Otol 1995 Sep;109(9):873-5
Colquitt WL, Zeh MC, Killian CD, Cultice JM. Effect of debt on U.S. medical school graduates' preferences for family medicine, general internal medicine, and general pediatrics. Acad Med 1996 Apr;71(4):399-411
Cols Jiménez M see **Tuneu Valls L**
Colson AM see **Brasseur G**
Colson AM see **Meunier B**
Colson C see **Hublet C**
Colson KL see **Zein N**
Colson P, Bailly C, Houssier C. Electric linear dichroism as a new tool to study sequence preference in drug binding to DNA. Biophys Chem 1996 Jan 16;58(1-2):125-40
Colson P, Damoiseaux P, Brisbois J, Duvivier E, Levecque P, Roger JM, Boulliez DJ, McKenna P, Clement J. Epidémie d'hantavirose dans l'Entre-Sambre-et-Meuse: année 1992-1993 Données cliniques et biologiques. Acta Clin Belg 1995;50(4):197-206 (Eng. Abstr.) **(Fre)**

FIGURA 66
Exemplo de entradas do índice de autores do *Cumulated Index Medicus*

Índices classificados

Existem basicamente dois tipos de índices classificados. Num deles, as entradas aparecem sob números de classificação altamente específicos extraídos de um esquema de classificação geral ou especializado. Este foi o método adotado pelo *Library and Information Science Abstracts* (LISA) até 1993. No *LISA* as entradas eram dispostas segundo um esquema de classificação facetada dedicado ao campo especializado da biblioteconomia e ciência da informação. A

figura 70 mostra algumas entradas relativas a cederrom. Observe-se como a notação relativa a bases de dados em formato de cederrom (*Zjjc*) é subdividida por meio de notações de outras partes da classificação (*Rn, Vtic*), a fim de oferecer maior especificidade, e como uma legenda textual é empregada para explicar cada notação específica. A figura 71 apresenta exemplos de entradas do índice alfabético de assuntos, inclusive algumas relativas aos itens mostrados na figura 70. Observe-se como os termos empregados como legendas textuais na figura 70 tornam-se pontos de entrada no índice de assuntos. O princípio adotado é o da indexação em cadeia (ver capítulo 4); cada nível da cadeia hierárquica é indexado a partir do mais específico até o mais genérico:

Magnetohydrodynamics
See also
Plasma (Physics)
Plasma waves
Synchrotron radiation
Alpha-torque forces. P. Graneau. bibl il diags *Electron Wirel World* 95:556-9 Je '89; Discussion. 95:875-6 S '89
Drop-on-demand operation of continuous jets using EHD techniques. D. W. Hrdina and J. M. Crowley. bibl flow chart diags *IEEE Trans Ind Appl* 25:705-10 Jl/Ag '89
Hydrodynamics of double-charged ions in a plane low-pressure discharge. D. A. Shapiro. bibl *J Phys D* 22:1107-13 Ag 14 '89
Iodine laser creates plasma X-rays. B. Dance. *Laser Focus World* 25:26+ Je '89
The magnetohydrodynamical instability of a current sheet created by plasma flow. A. I. Podgorny. bibl diags *Plasma Phys Control Fusion* 31:1271-9 Jl '89
A personal-computer-based package for interactive assessment of magnetohydrodynamic equilibrium and poloidal field coil design in axisymmetric toroidal geometry. W. P. Kelleher and D. Steiner. bibl diag *Fusion Technol* 15:1507-19 Jl '89
Why Extrap? B. Lehnert. bibl(p38-43) il diags *Fusion Technol* 16:7-43 Ag '89
Mathematical models
Induction electrohydrodynamic pump in a vertical configuration. J. Seyed-Yagoobi and others. bibl diags *J Heat Transf* 111:664-74 Ag '89
Mass transport and the bootstrap current from Ohm's law in steady-state tokamaks. J.-S. Kim and J. M. Greene. bibl *Plasma Phys Control Fusion* 31 no7:1069-94 Je '89
Reduction of thermal expansion in Z-pinches by electron beam assisted magnetic field generation. J. A. Heikkinen and S. J. Karttunen. bibl *Plasma Phys Control Fusion* 31 no7:1035-48 Je '89
Magnetometers
Scanner can detect brain damage. il *Engineer* 269:49 Ag 31-S 7 '89
Design
Electronic balancing of multichannel SQUID magnetometers. H. J. M. ter Brake and others. bibl diags *J Phys E* 22:560-4 Ag '89

FIGURA 67
Exemplo de entradas do *Applied Science and Technology Index*, 1986

graphite basal planes. A considerable fraction of the He gases desorbs at room temperature, implying that they are relatively mobile inside the lattice. (Author abstract) 32 Refs. English.

Choi, W. (Pohang Inst of Science and Technology, Pohang, SOUTH KOREA); Kim, C.; Kang, H. *Surf Sci* v 281 n 3 Feb 1 1993 p 323-335.

073651 **Light scattering as a probe of thermodynamic quantities in a binary mixture.** The authors have shown recently how Rayleigh-Brillouin light scattering can be used to extract certain thermodynamic quantities of a binary mixture in an approximate way. An approach which yields exact results is described here, although it requires knowledge of additional thermodynamic data. This information can be obtained either from other experiments or from a thermodynamic model prediction. Since we are dealing with a model system that can be described by a van der Waals equation of state, that model is preferred here. The results for the osmotic compressibility for a He + Xe mixture obtained in this way from the Landau-Placzek ratio are in good agreement with calculations. (Edited author abstract) 27 Refs. English.

Bot, Arjen (FOM, Amsterdam, Neth); Wegdam, Gerard H. *Fluid Phase Equilib* v 77 Sep 15 1992 p 285-295.

073652 **Role of the buffer gas in the ArF laser chemical vapour deposition of silicon oxide.** Inert gases are commonly used in thin film deposition methods as a diluent of the gas mixture or as a purging gas. However, several workers have determined the influence of the buffer gases on the film growth mechanism which has consequences for the film properties. In this paper, a study of the influence of argon, used as the reactor window purging gas, on the silicon oxide film growth and properties is presented. Films are deposited from silane and nitrous oxide by ArF-laser-induced chemical vapour deposition. By purging the beam entrance window with Ar, not only is window film formation or powder deposition avoided, but also reductions in the H and OH contents and thus better optical properties are achieved. (Author abstract) 28 Refs. English.

Gonzalez, P. (Univ of Vigo, Vigo, Spain); Pou, J.; Fernandez, D.; Garcia, E.; Serra, J.; Leon, B., Perez-Amor, M. *Thin Solid Films* v 230 n 1 Jul 15 1993 p 35-38.

INERTIAL CONFINEMENT FUSION

073653 **Analysis of radiation symmetrization in hohlraum targets.** Symmetrization of illumination non-uniformity by thermal radiation in spherical hohlraum targets has been studied systematically for indirectly driven inertial confinement fusion. Numerical calculations have shown that the effect of X-ray re-emission on the illumination uniformity is quasi-linear. On the basis of a linear theory it is found that the non-uniformity of each mode of the X-ray source considerably smoothed out by three different effects. The 1-st is the geometrical effect, which accounts for the configuration and the number of X-ray converters. The second is the effect of single emission, which depends only on the hohlraum structure (area ratio). The third is the effect of multiple re-emission, which is equal to the reciprocal of the average circulation number of radiation in a hohlraum target. The paper gives practical solutions regarding the required number of converters, N_c, in particular for heavy ion fusion systems. It is shown that $N_c = 6$ (hexahedron) is a necessary and sufficient condition to ensure tolerable symmetrical illumination (≤1-2% rms). (Author abstract) 19 Refs. English.

Murakami, M. (Inst for Laser Technology, Osaka, Jpn) *Nucl Fusion* v 32 n 10 Oct 1992 p 1715-1724.

073654 **Experimental testing of thin-shell stable acceleration for ICF schemes with direct and indirect drive.** The present review is of the experimental investigations on laser-plasma interaction being carried out in past years at IAE. Experiments were conducted on the 'Mishen' facility. The laser system of Mishen consists of two channels with output beam parameters as follows: the main beam-output energy 100-200 J (λ = 1.054 μm) in 3-ns pulse, divergence approx. 2×10^{-4} rad, contrast ratio approx. 10^4, power density at the target surface approx. 10^{13}-10^{14} W/cm²; the diagnostic beam - output energy 10-20 J (λ = 1.054 μm) and 5-10 J (λ = 0.53 μm) in 0.3-ns pulse, divergence approx. 10^{-3}-10^{-4} W/cm². Our aim in this experiment is to study the different aspects of the ICF processes in flat geometry. The main issues of our studies are hydrodynamic aspects, including acceleration efficiency, high-velocity impact in cascade targets, hydrostability, and X-ray physics-conversion efficiency, heat transfer, and X-ray-driven targets. Refs. English.

Bolotin, V.A. (Branch of Kurchatov Atomic Energy Inst, Moscow, Russia); Burdonsky, I.N.; Velikovich, A.L.; Gavrilov, V.V.; Golberg, S.M.; Goltsov, A.Yu.; Zhurkhukalo, E.V.; Zavjalets, S.V.; Kondrashov, V.N.; Kovalsky, N.G.; Pergament, M.I.; Koshelev, M.O.; Rupasov, A.A.; Shikanov, A.S. *Laser Part Beams* v 11 n 1 1993 Japan-US Symposium on Physics of High Power Laser Matter Interaction, Kyoto, Jpn, p 127-135.

073655 **Heavy-ion-driven targets for small-scale inertial confinement fusion experiments.** Two regimes of hydrodynamic evolution are found in the analysis of the performance of small-scale heavy-ion-driven targets. One leads to high density and high compression with moderate temperatures (approximately 1 keV) for driving energies of 100 kJ for 0.1-mg deuterium-tritium targets. Ignition can then be triggered by a second ion pulse (approximately 50 kJ). Breakeven could be obtained if a burnup fraction as small as 1% is obtained. The second regime leads to very high temperatures in the central part of the fuel, while the rest of the fuel remains at moderate temperatures (≤1 KeV), and the density is very low everywhere. Propagated ignition cannot occur in this case because of the small optical thickness of the compressed fuel (≤0.1 g/cm²). (Author abstract) 36 Refs. English.

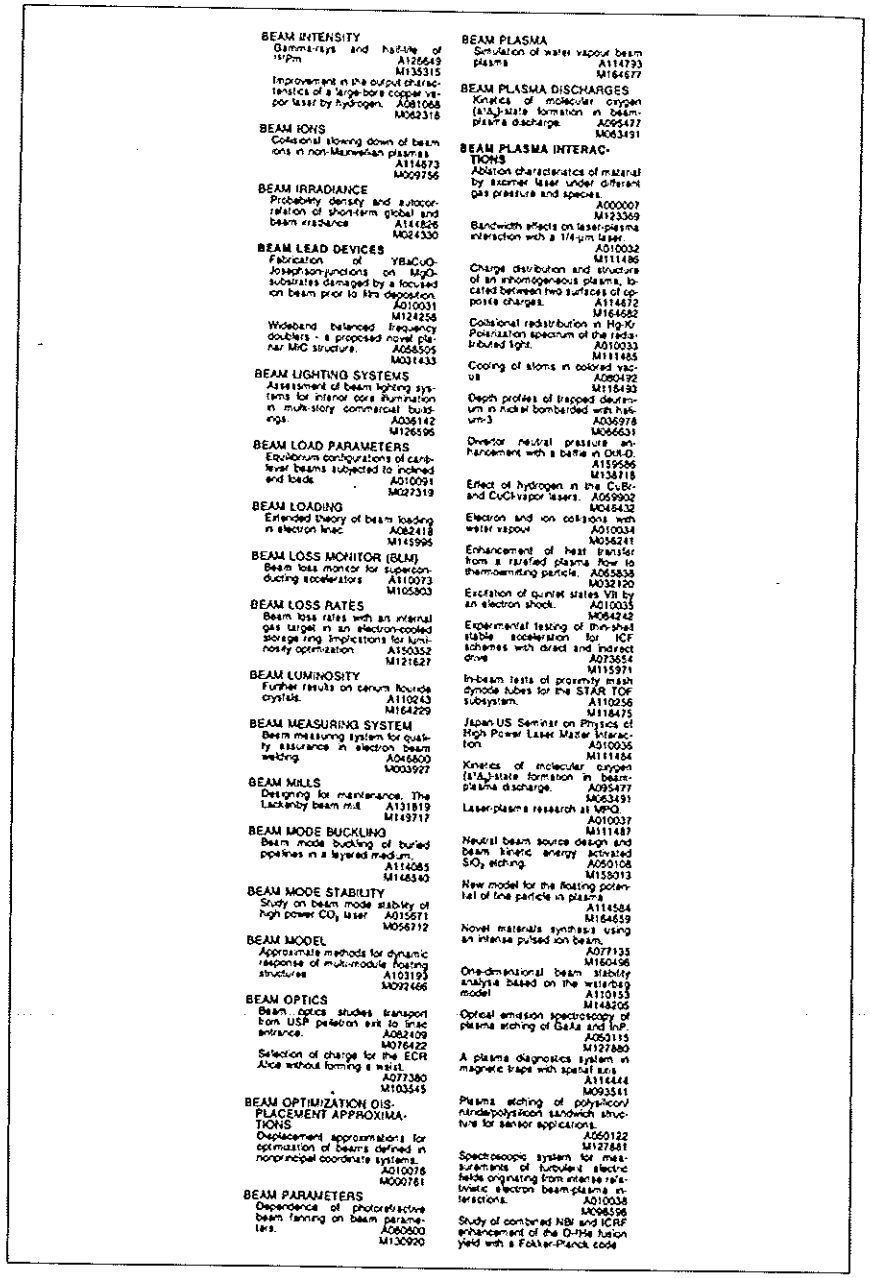
Martinez-Val, Jose M. (Madrid Polytechnic Univ, Madrid, Spain); Piera, Mireia. *Fusion Technol* v 23 n 2 Mar 1993 p 218-226.

073656 **High gain DT targets for heavy ion beam fusion.** In a parametric study of reactor size DT targets driven by beams of heavy ions it was found that spark ignition and high energy gains can be achieved in four-layer single-shell targets irradiated by a non-shaped box pulse of 10 GeV ^{209}Bi ions. With an input energy of E_{in}≈6 MJ delivered in t_p≤10 ns, one-dimensional energy gains of G≥400 are possible in the optimum cases. It is shown that, to obtain spark ignition and high energy gain, two conditions must be necessarily met: (1) a high enough implosion velocity, U_p≥6.2×10^7 $\Omega^{1/2}$ cm/s, must be reached, and (2) the fuel compression must be accomplished with a low enough pusher/fuel mass ratio, M_p/M_{DT}≤5-7 (Ω is a dimensionless parameter determined by the density distribution in the compressed target core). It was found also that when the (ρR) of the cold part of the compressed fuel is ≈2-5 g/cm², the main portion of the fuel is ignited owing to the heating by 14 MeV neutrons emitted from the central hot region. (Author abstract) 36 Refs. English.

Basko, M.M. (Max-Planck-Inst fur Quantenoptik, Munich, Ger). *Nucl Fusion* v 32 n 9 Sep 1992 p 1515-1529.

073657 **Low activation structural materials for ICF reactors: differences with MCF environments.** Activation calculations considering the neutron flux and spectrum of a first structural wall (FSW) in an inertial confinement fusion reactor (ICF) are performed for all stable elements, using a recently upgraded data base. Surface γ dose rate and waste disposal ratings (WDR) are employed as indices to compare the merit of elements and compute the concentration limits corresponding to hands-on processing, remote recycling and shallow land burial (SLB). The performance of steels, vanadium alloys and silicon carbide, as candidate structural materials has also been explored. The materials with less waste/recycling concerns are identified, and the influence that impurities

FIGURA 68
Exemplo de entradas do volume anual do *Engineering Index* (1993)



BEAM INTENSITY
Gamma-rays and half-life of 191Pm A126649 M133315
Improvement in the output characteristics of a large-bore copper vapor laser by hydrogen. A081068 M062318

BEAM IONS
Collisional slowing down of beam ions in non-Maxwellian plasmas A114673 M099756

BEAM IRRADIANCE
Probability density and autocorrelation of short-term global and beam irradiance A144826 M024330

BEAM LEAD DEVICES
Fabrication of YBaCuO-Josephson-junctions on MgO-substrates damaged by a focused ion beam prior to film deposition. A010031 M124258
Wideband balanced frequency doublers - a proposed novel planar MIC structure. A058505 M031433

BEAM LIGHTING SYSTEMS
Assessment of beam lighting systems for interior core illumination in multi-story commercial buildings. A036142 M126596

BEAM LOAD PARAMETERS
Equilibrium configurations of cantilever beams subjected to inclined end loads. A010091 M027319

BEAM LOADING
Extended theory of beam loading in electron linac A082418 M145996

BEAM LOSS MONITOR (BLM)
Beam loss monitor for superconducting accelerators A110073 M105803

BEAM LOSS RATES
Beam loss rates with an internal gas target in an electron-cooled storage ring: implications for luminosity optimization A150352 M121627

BEAM LUMINOSITY
Further results on cerium fluoride crystals. A110243 M164229

BEAM MEASURING SYSTEM
Beam measuring system for quality assurance in electron beam welding A046800 M003927

BEAM MILLS
Designing for maintenance. The Lackenby beam mill A131819 M149717

BEAM MODE BUCKLING
Beam mode buckling of buried pipelines in a layered medium. A114085 M146340

BEAM MODE STABILITY
Study on beam mode stability of high power CO_2 laser A015671 M056712

BEAM MODEL
Approximate methods for dynamic response of multi-module floating structures A103193 M092466

BEAM OPTICS
Beam optics studies transport from USP pelletron exit to linac entrance. A082409 M076422
Selection of charge for the ECR Xe+ without forming a metal. A077380 M103545

BEAM OPTIMIZATION DISPLACEMENT APPROXIMATIONS
Displacement approximations for optimization of beams defined in nonprincipal coordinate systems. A010076 M000781

BEAM PARAMETERS
Dependence of photorefractive beam fanning on beam parameters. A060600 M130920

BEAM PLASMA
Simulation of water vapour beam plasma A114793 M164677

BEAM PLASMA DISCHARGES
Kinetics of molecular oxygen ($a^1\Delta_g$)-state formation in beam-plasma discharge. A095477 M063491

BEAM PLASMA INTERACTIONS
Ablation characteristics of material by excimer laser under different gas pressure and species. A000007 M123369
Bandwidth effects on laser-plasma interaction with a 1/4-μm laser. A010032 M111486
Charge distribution and structure of an inhomogeneous plasma, located between two surfaces of opposite charges. A114672 M164682
Collisional redistribution in Hg-Kr. Polarization spectrum of the redistributed light. A010033 M111485
Cooling of atoms in collapsed vacua A080492 M116493
Depth profiles of trapped deuterium in nickel bombarded with helium-3 A036978 M066631
Director neutral pressure enhancement with a baffle in DIII-D. A155986 M138718
Effect of hydrogen in the CuBr- and CuCl-vapor lasers. A059902 M045432
Electron and ion collisions with water vapour A010034 M056241
Enhancement of heat transfer from a rarefied plasma flow to thermoemitting particle. A065838 M032120
Excitation of quintet states VII by an electron shock. A010035 M064242
Experimental testing of thin-shell stable acceleration for ICF schemes with direct and indirect drive A073654 M115971
In-beam tests of proximity mesh dynode tubes for the STAR TOF subsystem. A110256 M118475
Japan-US Seminar on Physics of High Power Laser Matter Interaction A010036 M111484
Kinetics of molecular oxygen ($a^1\Delta_g$)-state formation in beam-plasma discharge. A095477 M063491
Laser-plasma research at MPQ. A010037 M111481
Neutral beam source design and beam kinetic energy activated SiO_2 etching. A050108 M158013
New model for the floating potential of fine particle in plasma A114584 M164659
Novel materials synthesis using an intense pulsed ion beam. A077135 M160496
One-dimensional beam stability analysis based on the water-bag model A110153 M148205
Optical emission spectroscopy of plasma etching of GaAs and InP. A050115 M127880
A plasma diagnostics system in magnetic traps with spatial scan A114444 M093561
Plasma etching of polysilicon/nitride/polysilicon sandwich structure for sensor applications. A060122 M127881
Spectroscopic system for measurements of turbulent electric fields originating from energetic relativistic electron beam-plasma interactions. A010038 M096598
Study of combined NBI and ICRF enhancement of the D-3He fusion yield with a Fokker-Planck code

FIGURA 69
Exemplo de entradas do índice de assuntos do *Engineering Index* (1993)

ZjjcRaNak—CD-ROMs. Data bases. Information services.
Economic aspects **88/5339**
The CD-ROM marketplace: a producer's perspective. Christopher Pooley. *Wilson Library Bulletin*, 62 (4) Dec 87, 24-26.
Contribution to a special issue devoted in part to CD-ROM. When laser disc technology was first introduced to libraries, librarians recognised the great possibilities of the medium, especially its vast storage capacity. Examines the major differences between print, on-line and CD-ROM versions of the same data base which fall in 3 key areas: content, currency or update frequency, and pricing. Discusses competition in the marketplace and emphasises that the future for CD-ROM in libraries is excellent with more products, more new products offering combinations of data bases, better software and networking systems available to stimulate the growing use of CD-ROM products. (A.G.)

ZjjcRaNke—CD-ROMs. Data bases. Information services.
Cost-benefit analysis. PsycLIT **88/5340**
Justifying CD-ROM. Ralph Alberico. *Small Computers in Libraries*, 7 (2) Feb 87, 18-20.
Considers data bases on CD-ROM in terms of costs and benefits, by examining PsycLIT, an abbreviated CD-ROM version of the PsycINFO data base. PsycLIT was one of the first CD-ROM data bases and is one of the best and one of the most expensive. Compares the CD-ROM version to print and on-line products and stresses that as the number of users grow prices will decrease. (P.B.)

ZjjcVtlc—CD-ROMs. User-System interface **88/5341**
Entering unchartered territory: putting CD-ROM in place. Nancy Crane, Tamara Durfee. *Wilson Library Bulletin*, 62 (4) Dec 87, 28-30. illus.
Contribution to a special issue devoted in part to CD-ROM. Discusses considerations that need to be raised before implementing end-user CD-ROM and offers some proposals as to how solutions may be found. These include: assessing the environment; choice of a CD-ROM system; components needed in setting up a workstation; vendor services; placement; user constraints; training for searching; statistics and ongoing assessment; effects on staff; and desired features of CD-ROM services. The rationale for end-user CD-ROM is presented. (A.G.)

FIGURA 70
Exemplo de entradas do *Library and Information Science Abstracts* (antes de 1993)

Cost benefit analysis, Information services, Databases, CD-ROMs, Computerized information storage and retrieval

Information services, Databases, CD-ROMs, Computerized information storage and retrieval

Databases, CD-ROMs, Computerized information storage and retrieval

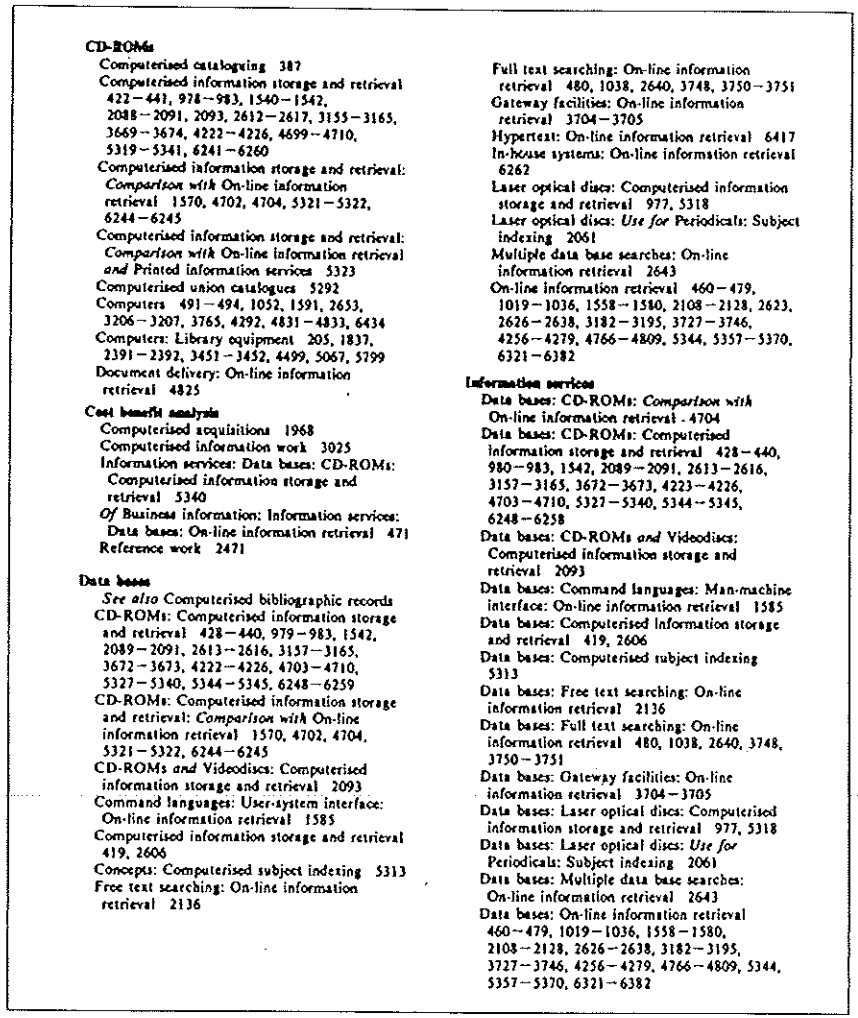
Computerized information storage and retrieval (esta entrada mais genérica não aparece na figura 71)

PsycLIT (o nome de uma base de dados) na figura 70 não foi um termo de indexação genuíno no *LISA* e, por isso, não deu origem a uma entrada no índice de assuntos, embora tenha originado uma entrada no índice de nomes próprios que é separado do de assuntos.

Enquanto o *LISA* empregava um esquema de classificação especializada, outros índices impressos se baseavam em esquemas gerais, dos quais a *Classificação Decimal Universal* (CDU) é o mais comumente adotado.

No outro método classificado utilizado na organização de uma base de dados impressa, as entradas são agrupadas sob categorias de assuntos relativa-

mente genéricas, proporcionando-se acesso a assuntos mais específicos por meio de índices. Um exemplo é o *LISA* atual. A figura 72 mostra as categorias genéricas de assuntos sob as quais os resumos foram organizados a partir de 1997, e a figura 73 mostra exemplos de algumas entradas. O índice de assuntos ainda se baseia em processos de indexação em cadeia (ver figura 74) embora não mais estejam atrelados a um esquema de classificação.



CD-ROMs
Computerised cataloguing 387
Computerised information storage and retrieval 422–441, 978–983, 1540–1542, 2088–2091, 2093, 2612–2617, 3155–3165, 3669–3674, 4222–4226, 4699–4710, 5319–5341, 6241–6260
Computerised information storage and retrieval: *Comparison with* On-line information retrieval 1570, 4702, 4704, 5321–5322, 6244–6245
Computerised information storage and retrieval: *Comparison with* On-line information retrieval *and* Printed information services 5323
Computerised union catalogues 5292
Computers 491–494, 1052, 1591, 2653, 3206–3207, 3765, 4292, 4831–4833, 6434
Computers: Library equipment 205, 1837, 2391–2392, 3451–3452, 4499, 5067, 5799
Document delivery: On-line information retrieval 4825

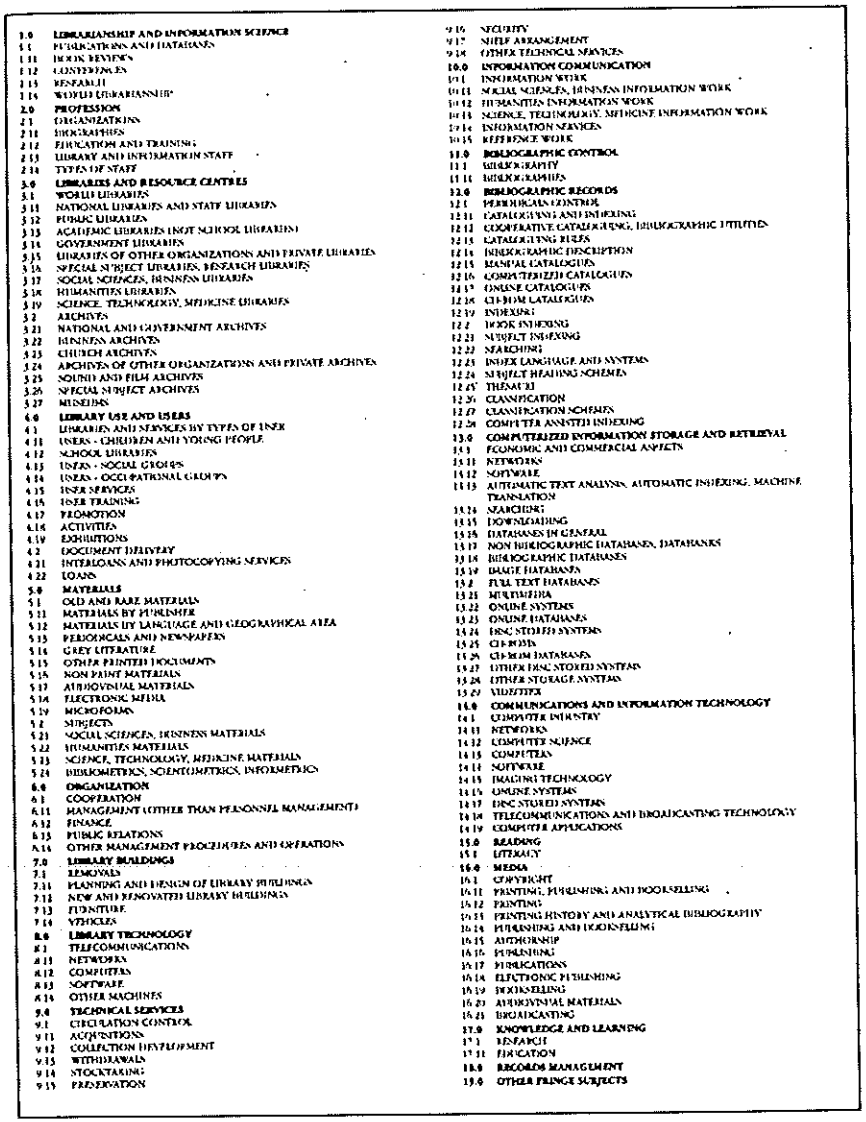
Cost benefit analysis
Computerised acquisitions 1968
Computerised information work 3025
Information services: Data bases: CD-ROMs: Computerised information storage and retrieval 5340
Of Business information: Information services: Data bases: On-line information retrieval 471
Reference work 2471

Data bases
See also Computerised bibliographic records
CD-ROMs: Computerised information storage and retrieval 428–440, 979–983, 1542, 2089–2091, 2613–2616, 3157–3165, 3672–3673, 4222–4226, 4703–4710, 5327–5340, 5344–5345, 6248–6259
CD-ROMs: Computerised information storage and retrieval: *Comparison with* On-line information retrieval 1570, 4702, 4704, 5321–5322, 6244–6245
CD-ROMs *and* Videodiscs: Computerised information storage and retrieval 2093
Command languages: User-system interface: On-line information retrieval 1585
Computerised information storage and retrieval 419, 2606
Concepts: Computerised subject indexing 5313
Free text searching: On-line information retrieval 2136
Full text searching: On-line information retrieval 480, 1038, 2640, 3748, 3750–3751
Gateway facilities: On-line information retrieval 3704–3705
Hypertext: On-line information retrieval 6417
In-house systems: On-line information retrieval 6262
Laser optical discs: Computerised information storage and retrieval 977, 5318
Laser optical discs: *Use for* Periodicals: Subject indexing 2061
Multiple data base searches: On-line information retrieval 2643
On-line information retrieval 460–479, 1019–1036, 1558–1580, 2108–2128, 2623, 2626–2638, 3182–3195, 3727–3746, 4256–4279, 4766–4809, 5344, 5357–5370, 6321–6382

Information services
Data bases: CD-ROMs: *Comparison with* On-line information retrieval 4704
Data bases: CD-ROMs: Computerised information storage and retrieval 428–440, 980–983, 1542, 2089–2091, 2613–2616, 3157–3165, 3672–3673, 4223–4226, 4703–4710, 5327–5340, 5344–5345, 6248–6258
Data bases: CD-ROMs *and* Videodiscs: Computerised information storage and retrieval 2093
Data bases: Command languages: Man-machine interface: On-line information retrieval 1585
Data bases: Computerised information storage and retrieval 419, 2606
Data bases: Computerised subject indexing 5313
Data bases: Free text searching: On-line information retrieval 2136
Data bases: Full text searching: On-line information retrieval 480, 1038, 2640, 3748, 3750–3751
Data bases: Gateway facilities: On-line information retrieval 3704–3705
Data bases: Laser optical discs: Computerised information storage and retrieval 977, 5318
Data bases: Laser optical discs: *Use for* Periodicals: Subject indexing 2061
Data bases: Multiple data base searches: On-line information retrieval 2643
Data bases: On-line information retrieval 460–479, 1019–1036, 1558–1580, 2108–2128, 2626–2638, 3182–3195, 3727–3746, 4256–4279, 4766–4809, 5344, 5357–5370, 6321–6382

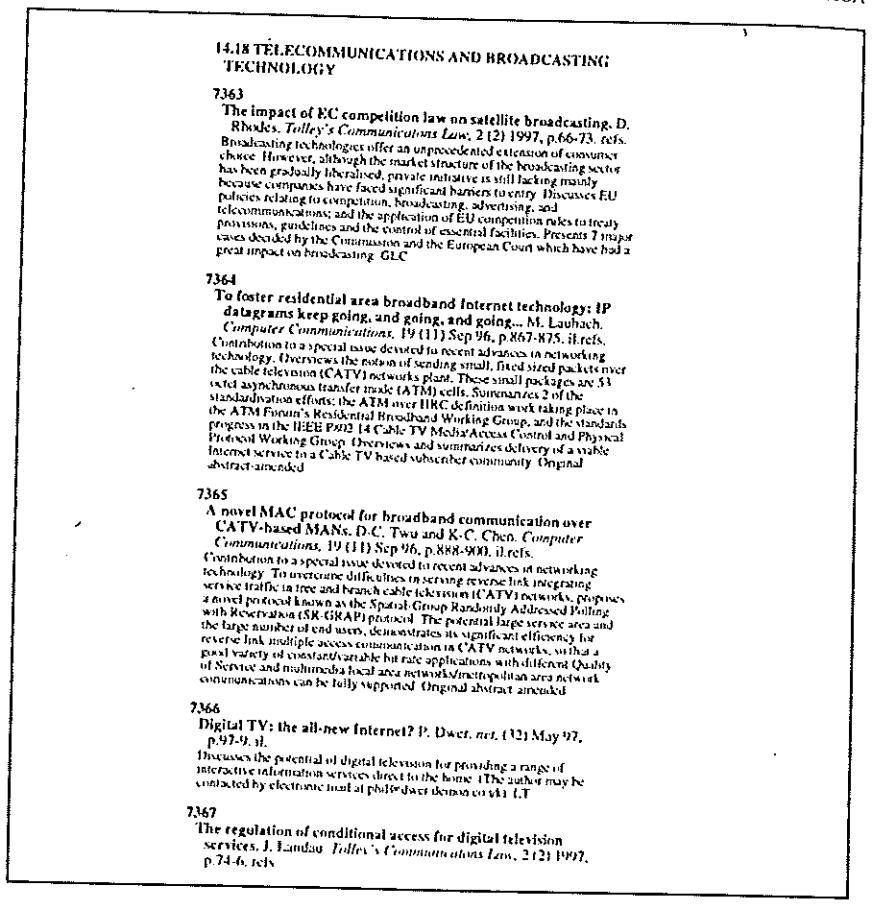
FIGURA 71
Exemplo de entradas do índice de assuntos do *Library and Information Science Abstracts* (antes de 1993)
Reproduzido com permissão do editor

O *Chemical Abstracts* assemelha-se ao *LISA* porque as entradas são organizadas sob categorias e subcategorias temáticas. O índice de assuntos, no entanto, é bastante diferente, estando baseado no princípio de articulação (ver capítulo 4): cadeias de termos atribuídos por indexadores humanos são manipuladas



1.0 LIBRARIANSHIP AND INFORMATION SCIENCE
1.1 PUBLICATIONS AND DATABASES
1.11 BOOK REVIEWS
1.12 CONFERENCES
1.13 RENFAID
1.14 WORLD LIBRARIANSHIP
2.0 PROFESSION
2.1 ORGANIZATIONS
2.11 BIOGRAPHIES
2.12 EDUCATION AND TRAINING
2.13 LIBRARY AND INFORMATION STAFF
2.14 TYPES OF STAFF
3.0 LIBRARIES AND RESOURCE CENTRES
3.1 WORLD LIBRARIES
3.11 NATIONAL LIBRARIES AND STATE LIBRARIES
3.12 PUBLIC LIBRARIES
3.13 ACADEMIC LIBRARIES (NOT SCHOOL LIBRARIES)
3.14 GOVERNMENT LIBRARIES
3.15 LIBRARIES OF OTHER ORGANIZATIONS AND PRIVATE LIBRARIES
3.16 SPECIAL SUBJECT LIBRARIES, RESEARCH LIBRARIES
3.17 SOCIAL SCIENCES, BUSINESS LIBRARIES
3.18 HUMANITIES LIBRARIES
3.19 SCIENCE, TECHNOLOGY, MEDICINE LIBRARIES
3.2 ARCHIVES
3.21 NATIONAL AND GOVERNMENT ARCHIVES
3.22 BUSINESS ARCHIVES
3.23 CHURCH ARCHIVES
3.24 ARCHIVES OF OTHER ORGANIZATIONS AND PRIVATE ARCHIVES
3.25 SOUND AND FILM ARCHIVES
3.26 SPECIAL SUBJECT ARCHIVES
3.27 MUSEUMS
4.0 LIBRARY USE AND USERS
4.1 LIBRARIES AND SERVICES BY TYPES OF USER
4.11 USERS - CHILDREN AND YOUNG PEOPLE
4.12 SCHOOL LIBRARIES
4.13 USERS - SOCIAL GROUPS
4.14 USERS - OCCUPATIONAL GROUPS
4.15 USER SERVICES
4.16 USER TRAINING
4.17 PROMOTION
4.18 ACTIVITIES
4.19 EXHIBITIONS
4.2 DOCUMENT DELIVERY
4.21 INTERLOANS AND PHOTOCOPYING SERVICES
4.22 LOANS
5.0 MATERIALS
5.1 OLD AND RARE MATERIALS
5.11 MATERIALS BY PUBLISHER
5.12 MATERIALS BY LANGUAGE AND GEOGRAPHICAL AREA
5.13 PERIODICALS AND NEWSPAPERS
5.14 GREY LITERATURE
5.15 OTHER PRINTED DOCUMENTS
5.16 NON PRINT MATERIALS
5.17 AUDIOVISUAL MATERIALS
5.18 ELECTRONIC MEDIA
5.19 MICROFORMS
5.2 SUBJECTS
5.21 SOCIAL SCIENCES, BUSINESS MATERIALS
5.22 HUMANITIES MATERIALS
5.23 SCIENCE, TECHNOLOGY, MEDICINE MATERIALS
5.24 BIBLIOMETRICS, SCIENTOMETRICS, INFORMETRICS
6.0 ORGANIZATION
6.1 COOPERATION
6.11 MANAGEMENT (OTHER THAN PERSONNEL MANAGEMENT)
6.12 FINANCE
6.13 PUBLIC RELATIONS
6.14 OTHER MANAGEMENT PROCEDURES AND OPERATIONS
7.0 LIBRARY BUILDINGS
7.1 REMOVALS
7.11 PLANNING AND DESIGN OF LIBRARY BUILDINGS
7.12 NEW AND RENOVATED LIBRARY BUILDINGS
7.13 FURNITURE
7.14 VEHICLES
8.0 LIBRARY TECHNOLOGY
8.1 TELECOMMUNICATIONS
8.11 NETWORKS
8.12 COMPUTERS
8.13 SOFTWARE
8.14 OTHER MACHINES
9.0 TECHNICAL SERVICES
9.1 CIRCULATION CONTROL
9.11 ACQUISITIONS
9.12 COLLECTION DEVELOPMENT
9.13 WITHDRAWALS
9.14 STOCKTAKING
9.15 PRESERVATION
9.16 SECURITY
9.17 SHELF ARRANGEMENT
9.18 OTHER TECHNICAL SERVICES
10.0 INFORMATION COMMUNICATION
10.1 INFORMATION WORK
10.11 SOCIAL SCIENCES, BUSINESS INFORMATION WORK
10.12 HUMANITIES INFORMATION WORK
10.13 SCIENCE, TECHNOLOGY, MEDICINE INFORMATION WORK
10.14 INFORMATION SERVICES
10.15 REFERENCE WORK
11.0 BIBLIOGRAPHIC CONTROL
11.1 BIBLIOGRAPHY
11.11 BIBLIOGRAPHIES
12.0 BIBLIOGRAPHIC RECORDS
12.1 PERIODICALS CONTROL
12.11 CATALOGUING AND INDEXING
12.12 COOPERATIVE CATALOGUING, BIBLIOGRAPHIC UTILITIES
12.13 CATALOGUING RULES
12.14 BIBLIOGRAPHIC DESCRIPTION
12.15 MANUAL CATALOGUES
12.16 COMPUTERIZED CATALOGUES
12.17 ONLINE CATALOGUES
12.18 CD-ROM CATALOGUES
12.19 INDEXING
12.2 BOOK INDEXING
12.21 SUBJECT INDEXING
12.22 SEARCHING
12.23 INDEX LANGUAGE AND SYSTEMS
12.24 SUBJECT HEADING SCHEMES
12.25 THESAURI
12.26 CLASSIFICATION
12.27 CLASSIFICATION SCHEMES
12.28 COMPUTER ASSISTED INDEXING
13.0 COMPUTERIZED INFORMATION STORAGE AND RETRIEVAL
13.1 ECONOMIC AND COMMERCIAL ASPECTS
13.11 NETWORKS
13.12 SOFTWARE
13.13 AUTOMATIC TEXT ANALYSIS, AUTOMATIC INDEXING, MACHINE TRANSLATION
13.14 SEARCHING
13.15 DOWNLOADING
13.16 DATABASES IN GENERAL
13.17 NON BIBLIOGRAPHIC DATABASES, DATABANKS
13.18 BIBLIOGRAPHIC DATABASES
13.19 IMAGE DATABASES
13.2 FULL TEXT DATABASES
13.21 MULTIMEDIA
13.22 ONLINE SYSTEMS
13.23 ONLINE DATABASES
13.24 DISC STORED SYSTEMS
13.25 CD-ROMS
13.26 CD-ROM DATABASES
13.27 OTHER DISC STORED SYSTEMS
13.28 OTHER STORAGE SYSTEMS
13.29 VIDEOTEX
14.0 COMMUNICATIONS AND INFORMATION TECHNOLOGY
14.1 COMPUTER INDUSTRY
14.11 NETWORKS
14.12 COMPUTER SCIENCE
14.13 COMPUTERS
14.14 SOFTWARE
14.15 IMAGING TECHNOLOGY
14.16 ONLINE SYSTEMS
14.17 DISC STORED SYSTEMS
14.18 TELECOMMUNICATIONS AND BROADCASTING TECHNOLOGY
14.19 COMPUTER APPLICATIONS
15.0 READING
15.1 LITERACY
16.0 MEDIA
16.1 COPYRIGHT
16.11 PRINTING, PUBLISHING AND BOOKSELLING
16.12 PRINTING
16.13 PRINTING HISTORY AND ANALYTICAL BIBLIOGRAPHY
16.14 PUBLISHING AND BOOKSELLING
16.15 AUTHORSHIP
16.16 PUBLISHING
16.17 PUBLICATIONS
16.18 ELECTRONIC PUBLISHING
16.19 BOOKSELLING
16.21 AUDIOVISUAL MATERIALS
16.23 BROADCASTING
17.0 KNOWLEDGE AND LEARNING
17.1 RENFAID
17.11 EDUCATION
18.0 RECORDS MANAGEMENT
19.0 OTHER FRINGE SUBJECTS

FIGURA 72
Categorias de assuntos usadas pelo *Library and Information Science Abstracts* (1997)
Reproduzido com permissão do editor



14.18 TELECOMMUNICATIONS AND BROADCASTING: TECHNOLOGY

7363
The impact of EC competition law on satellite broadcasting. D. Rhodes. *Tolley's Communications Law*; 2 (2) 1997, p.66-73. refs.
Broadcasting technologies offer an unprecedented extension of consumer choice. However, although the market structure of the broadcasting sector has been gradually liberalised, private initiative is still lacking mainly because companies have faced significant barriers to entry. Discusses EU policies relating to competition, broadcasting, advertising, and telecommunications; and the application of EU competition rules to treaty provisions, guidelines and the control of essential facilities. Presents 7 major cases decided by the Commission and the European Court which have had a great impact on broadcasting. GLC

7364
To foster residential area broadband Internet technology: IP datagrams keep going, and going, and going... M. Laubach. *Computer Communications*, 19 (11) Sep 96, p.867-875, il.refs.
Contribution to a special issue devoted to recent advances in networking technology. Overviews the notion of sending small, fixed sized packets over the cable television (CATV) networks plant. These small packages are 53 octet asynchronous transfer mode (ATM) cells. Summarizes 2 of the standardization efforts: the ATM over HRC definition work taking place in the ATM Forum's Residential Broadband Working Group, and the standards progress in the IEEE P802.14 Cable TV Media Access Control and Physical Protocol Working Group. Overviews and summarizes delivery of a viable Internet service to a Cable TV based subscriber community. Original abstract amended

7365
A novel MAC protocol for broadband communication over CATV-based MANs. D.C. Twu and K.C. Chen. *Computer Communications*, 19 (11) Sep 96, p.888-900, il.refs.
Contribution to a special issue devoted to recent advances in networking technology. To overcome difficulties in serving reverse link integrating service traffic in tree and branch cable television (CATV) networks, proposes a novel protocol known as the Spatial Group Randomly Addressed Polling with Reservation (SR-GRAP) protocol. The potential large service area and the large number of end users, demonstrates its significant efficiency for reverse link multiple access communication in CATV networks, so that a good variety of constant/variable bit rate applications with different Quality of Service and multimedia local area networks/metropolitan area network communications can be fully supported. Original abstract amended

7366
Digital TV: the all-new Internet? P. Dwet. *net*, (32) May 97, p.97-9, il.
Discusses the potential of digital television for providing a range of interactive information services direct to the home. (The author may be contacted by electronic mail at philf@wt.bemon.co.uk) LT

7367
The regulation of conditional access for digital television services. J. Landau. *Tolley's Communications Law*, 2 (2) 1997, p.74-6. refs.

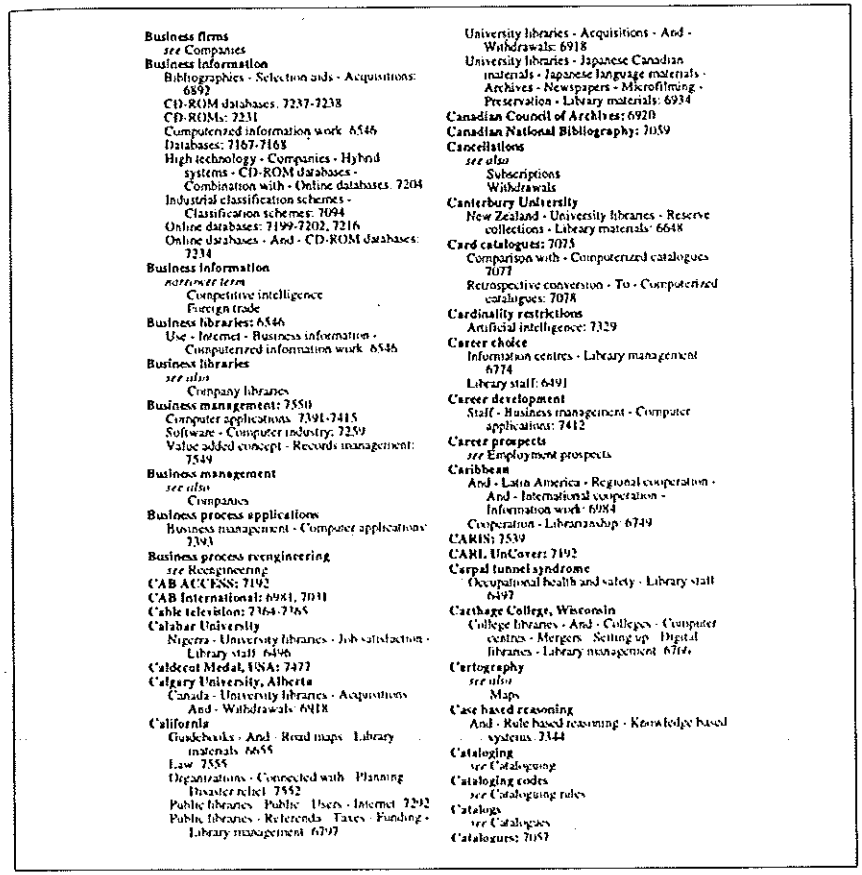
FIGURA 73
Exemplo de entradas do *Library and Information Science Abstracts*
Reproduzido com permissão do editor

de forma padronizada de modo a proporcionar um grupo de pontos de acesso coerentes para cada item (figura 75). Embora esse índice articulado de assuntos apareça somente nas acumulações do *Chemical Abstracts*, em cada fascículo semanal é publicado um índice de palavras-chave (ver figura 76). O *Chemical Abstracts* também inclui um índice de fórmulas químicas (ver figura 77).

Outros índices

A maioria dos outros serviços de indexação/resumos em formato impresso são variações dos tipos já exemplificados. O *Sociology of Education Abstracts*, diferentemente do *Library and Information Science Abstracts* e do *Chemical Abstracts*, simplesmente lista os resumos em ordem numérica sem agrupá-los

sob categorias genéricas de assuntos. O índice de assuntos, descrito como um 'índice de palavras-chave modificado', indexa os resumos sob palavras-chave ou expressões que aparecem no título ou no próprio resumo. Também são indexados os nomes próprios. A figura 78 mostra exemplos de dois resumos, e a figura 79 apresenta exemplo de entradas de índice, inclusive algumas correspondentes aos resumos da figura 78 (por exemplo, black dropouts [evasão escolar entre negros], class cutting [cábula na escola], compulsory education [ensino compulsório]).

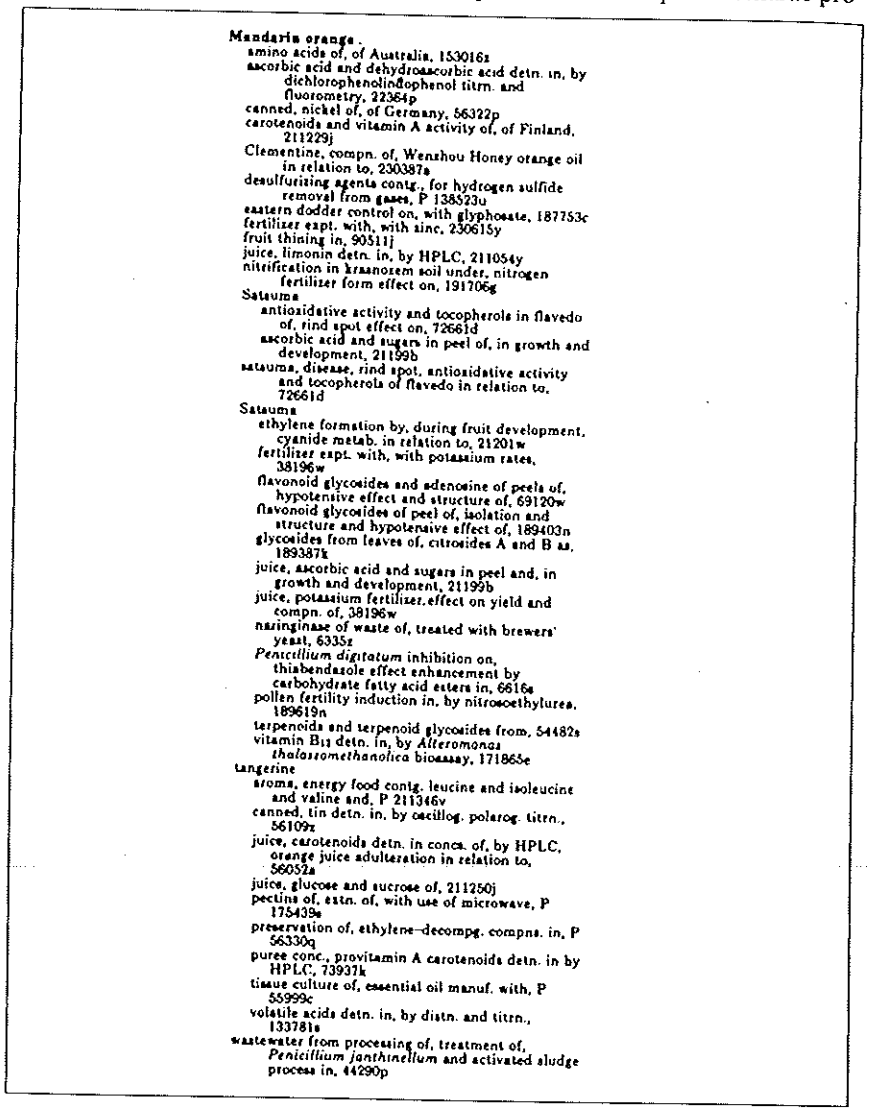


Business firms
see Companies
Business information
Bibliographies - Selection aids - Acquisitions: 6892
CD-ROM databases: 7237-7238
CD-ROMs: 7231
Computerized information work: 6546
Databases: 7167-7168
High technology - Companies - Hybrid systems - CD-ROM databases - Combination with - Online databases: 7204
Industrial classification schemes - Classification schemes: 7094
Online databases: 7199-7202, 7216
Online databases - And - CD-ROM databases: 7234
Business information
narrower term
Competitive intelligence
Foreign trade
Business libraries: 6546
Use - Internet - Business information - Computerized information work: 6546
Business libraries
see also
Company libraries
Business management: 7550
Computer applications: 7391-7415
Software - Computer industry: 7259
Value added concept - Records management: 7549
Business management
see also
Companies
Business process applications
Business management - Computer applications: 7393
Business process reengineering
see Reengineering
CAB ACCESS: 7192
CAB International: 6931, 7031
Cable television: 7364-7365
Calabar University
Nigeria - University libraries - Job satisfaction - Library staff: 6496
Caldecot Medal, USA: 7477
Calgary University, Alberta
Canada - University libraries - Acquisitions - And - Withdrawals: 6918
California
Guidebooks - And - Road maps - Library materials: 6655
Law: 7555
Organizations - Connected with - Planning - Disaster relief: 7552
Public libraries - Public - Users - Internet: 7292
Public libraries - Referenda - Taxes - Funding - Library management: 6797
University libraries - Acquisitions - And - Withdrawals: 6918
University libraries - Japanese Canadian materials - Japanese language materials - Archives - Newspapers - Microfilming - Preservation - Library materials: 6934
Canadian Council of Archives: 6920
Canadian National Bibliography: 7039
Cancellations
see also
Subscriptions
Withdrawals
Canterbury University
New Zealand - University libraries - Reserve collections - Library materials: 6648
Card catalogues: 7075
Comparison with - Computerized catalogues: 7077
Retrospective conversion - To - Computerized catalogues: 7078
Cardinality restrictions
Artificial intelligence: 7329
Career choice
Information centres - Library management: 6774
Library staff: 6491
Career development
Staff - Business management - Computer applications: 7412
Career prospects
see Employment prospects
Caribbean
And - Latin America - Regional cooperation - And - International cooperation - Information work: 6984
Cooperation - Librarianship: 6749
CARIS: 7539
CARL UnCover: 7192
Carpal tunnel syndrome
Occupational health and safety - Library staff: 6497
Carthage College, Wisconsin
College libraries - And - Colleges - Computer centres - Mergers - Setting up - Digital libraries - Library management: 6766
Cartography
see also
Maps
Case based reasoning
And - Rule based reasoning - Knowledge based systems: 7344
Cataloging
see Cataloguing
Cataloging codes
see Cataloguing rules
Catalogs
see Catalogues
Catalogues: 7057

FIGURA 74
Exemplo de entradas do índice de assuntos do *Library and Information Science Abstracts*
Reproduzido com permissão do editor

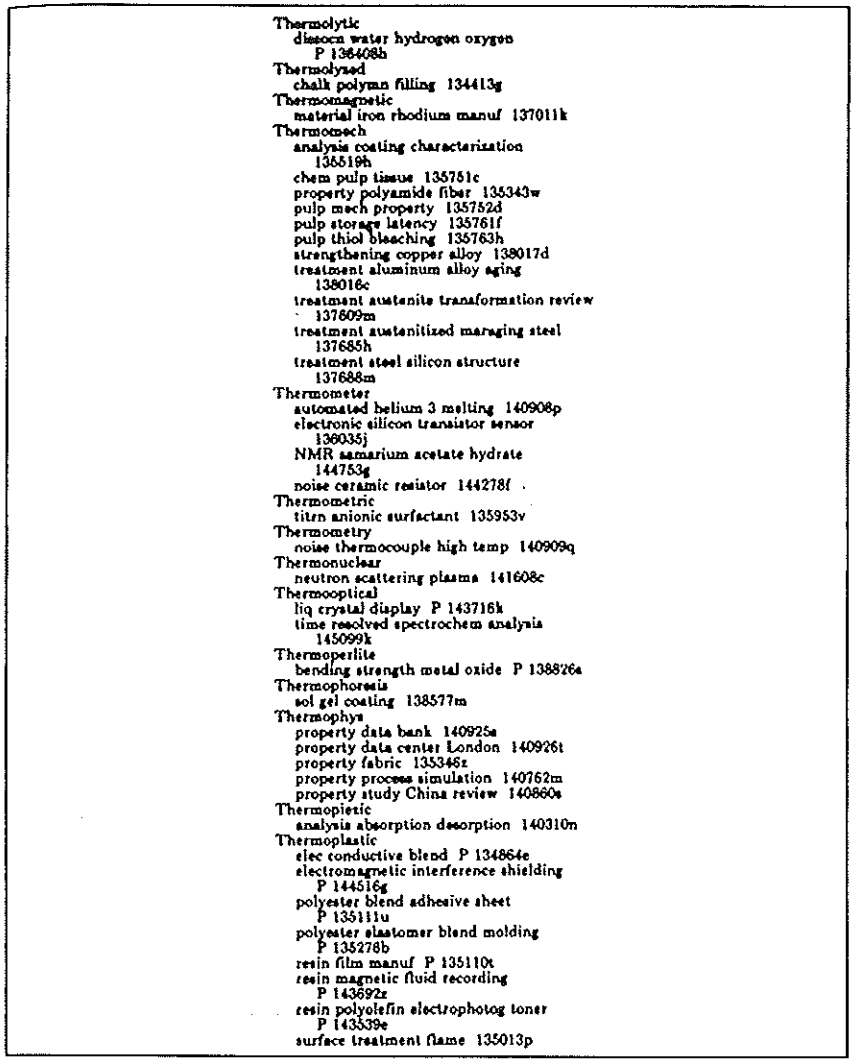
As inúmeras revistas de resumos publicadas na série da *Excerpta Medica* (Elsevier Science Publishers) também agrupam os itens sob categorias genéri-

cas de assuntos. Os índices de assuntos são altamente específicos. Todos os termos atribuídos (extraídos de um tesauro) por indexadores aparecem em cada entrada do índice. A maior parte desses termos tornam-se pontos de entrada no índice, sendo os outros termos mantidos como modificadores. Os modificadores são ordenados alfabeticamente em duas seqüências: termos que se tornarão pro-



Mandarin orange
amino acids of, of Australia, 153016z
ascorbic acid and dehydroascorbic acid detn. in, by dichlorophenolindophenol titrn. and fluorometry, 22364p
canned, nickel of, of Germany, 56322p
carotenoids and vitamin A activity of, of Finland, 211229j
Clementine, compn. of, Wenzhou Honey orange oil in relation to, 230387a
desulfurizing agents contg., for hydrogen sulfide removal from gases, P 138523u
eastern dodder control on, with glyphosate, 187753c
fertilizer expt. with, with zinc, 230615y
fruit thining in, 90511j
juice, limonin detn. in, by HPLC, 211054y
nitrification in kraanozem soil under, nitrogen fertilizer form effect on, 191706g
Satsuma
antioxidative activity and tocopherols in flavedo of, rind spot effect on, 72661d
ascorbic acid and sugars in peel of, in growth and development, 21199b
satsuma, disease, rind spot, antioxidative activity and tocopherols of flavedo in relation to, 72661d
Satsuma
ethylene formation by, during fruit development, cyanide metab. in relation to, 21201w
fertilizer expt. with, with potassium rates, 38196w
flavonoid glycosides and adenosine of peels of, hypotensive effect and structure of, 69120w
flavonoid glycosides of peel of, isolation and structure and hypotensive effect of, 189403n
glycosides from leaves of, citrosides A and B as, 189387k
juice, ascorbic acid and sugars in peel and, in growth and development, 21199b
juice, potassium fertilizer effect on yield and compn. of, 38196w
naringinase of waste of, treated with brewers' yeast, 6335z
Penicillium digitatum inhibition on, thiabendazole effect enhancement by carbohydrate fatty acid esters in, 6616a
pollen fertility induction in, by nitrosoethylurea, 189619n
terpenoids and terpenoid glycosides from, 54482a
vitamin B_{12} detn. in, by *Alteromonas thalassomethanolica* bioassay, 171865e
tangerine
aroma, energy food contg. leucine and isoleucine and valine and, P 211346v
canned, tin detn. in, by oscillog. polarog. titrn., 56109z
juice, carotenoids detn. in concn. of, by HPLC, orange juice adulteration in relation to, 56052a
juice, glucose and sucrose of, 211250j
pectins of, extn. of, with use of microwave, P 175439a
preservation of, ethylene–decompg. compns. in, P 56330q
puree conc., provitamin A carotenoids detn. in by HPLC, 73937k
tissue culture of, essential oil manuf. with, P 55999c
volatile acids detn. in, by distn. and titrn., 133781a
wastewater from processing of, treatment of, *Penicillium janthinellum* and activated sludge process in, 44290p

FIGURA 75
Exemplo de entradas do índice de assuntos do *Chemical Abstracts*
Reproduzido com permissão do Chemical Abstracts Service



Thermolytic
dissocn water hydrogen oxygen P 136408h
Thermolyzed
chalk polymn filling 134413g
Thermomagnetic
material iron rhodium manuf 137011k
Thermomech
analysis coating characterization 136519h
chem pulp tissue 135751c
property polyamide fiber 135343w
pulp mech property 135752d
pulp storage latency 135761f
pulp thiol bleaching 135763h
strengthening copper alloy 138017d
treatment aluminum alloy aging 138016c
treatment austenite transformation review 137609m
treatment austenitized maraging steel 137685h
treatment steel silicon structure 137688m
Thermometer
automated helium 3 melting 140908p
electronic silicon transistor sensor 136035j
NMR samarium acetate hydrate 144753g
noise ceramic resistor 144278f
Thermometric
titrn anionic surfactant 135953v
Thermometry
noise thermocouple high temp 140909q
Thermonuclear
neutron scattering plasma 141608c
Thermooptical
liq crystal display P 143716k
time resolved spectrochem analysis 145099k
Thermoperlite
bending strength metal oxide P 138826a
Thermophoresis
sol gel coating 138577m
Thermophys
property data bank 140925a
property data center London 140926t
property fabric 135346z
property process simulation 140762m
property study China review 140860s
Thermopietic
analysis absorption desorption 140310n
Thermoplastic
elec conductive blend P 134864e
electromagnetic interference shielding P 144516g
polyester blend adhesive sheet P 135111u
polyester elastomer blend molding P 135278b
resin film manuf P 135110t
resin magnetic fluid recording P 143692r
resin polyolefin electrophotog toner P 143539e
surface treatment flame 135013p

FIGURA 76
Exemplo de entradas do índice de palavras-chave do *Chemical Abstracts*
Reproduzido com permissão do Chemical Abstracts Service

priamente pontos de entrada precedem os termos que são apenas modificadores e não servirão como pontos de entrada. A figura 80 mostra um exemplo disso. Observe-se como a cadeia de termos funciona como uma espécie de minirresumo, oferecendo uma clara indicação (na maioria dos casos) daquilo de que trata cada item. Os índices de assuntos da *Excerpta Medica* são examinados com mais detalhes no capítulo 4.

A maioria dos índices alfabético-específicos organiza as referências bibliográficas sob cabeçalhos de assuntos, às vezes com subcabeçalhos, e entradas repetidas sob dois ou mais cabeçalhos (como no *Index Medicus*), ou organizam os resumos sob cabeçalhos de assuntos e adotam alguma forma de índice que proporciona possibilidades alternativas de acesso por assunto a itens isolados (como no *Engineering Index*). Há variações deste método alfabético-específico.

$C_{52}H_{44}N_8O_2P_2Tc$
Technetium, [1,3-bis(4-methylphenyl)-1-=
triazenato-N^1,N^3]dicarbonylbis=
(triphenylphosphine)-
(*OC*-6-14)- *[99354-95-7]*, 14057b

$C_{52}H_{44}N_4O_6P_2$
Phosphonic acid, (3,3',4,4',6,6'-hexaphenyl[6,6'-=
bi-6*H*-pyrrolo[1,2-*b*]pyrazole]-2,2'-diyl)bis-
tetramethyl ester, (*R**,*S**)- *[100418-78-8]*,
88671u
tetramethyl ester, (*R**,*S**)-, compd. with
trichloromethane (1:1), monohydrate
[100418-79-9], 88671u

$C_{52}H_{44}N_4OZn$
Zinc, [4-(diethylamino)-*N*-[2-(10,15,20-=
triphenyl-21*H*,23*H*-porphin-5-yl)phenyl]=
butanamidato(2-)-N^{21},N^{22},N^{23},N^{24}]-
(*SP*-4-2)- *[102497-59-6]*, 224763e

$C_{52}H_{44}N_6O_4$
2-Naphthalenecarboxamide, 4,4'-[(3,3',5,5'-=
tetramethyl[1,1'-biphenyl]-4,4'-diyl)bis=
(azo)]bis[3-hydroxy-*N*-(4-methylphenyl)-
[81287-27-6], P 196932p

$C_{52}H_{44}N_6O_6$
2-Naphthalenecarboxamide, 4,4'-[(3,3',5,5'-=
tetramethyl[1,1'-biphenyl]-4,4'-diyl)bis=
(azo)]bis[3-hydroxy-*N*-(2-methoxyphenyl)-
[81287-28-7], P 196932p

$C_{52}H_{44}N_8O_4$
2-Naphthalenecarboxamide, 4,4'-[1,4-piperazine=
diylbis(4,1-phenyleneazo)]bis[3-hydroxy-=
N-(4-methylphenyl)- *[101701-09-1]*, P
196932p

$C_{52}H_{44}N_8O_6$
2-Naphthalenecarboxamide, 4,4'-[1,4-piperazine=
diylbis(4,1-phenyleneazo)]bis[3-hydroxy-=
N-(3-methoxyphenyl)- *[101701-10-4]*, P
196932p

$C_{52}H_{44}O_2Sb_2$
Antimony, bis(benzeneacetato-*O*)-μ-=
oxohexaphenyldi-
stereoisomer *[99825-05-5]*, 50926t

$C_{52}H_{44}P_2Rh$
Rhodium(1+), [[1,1'-binaphthalene]-2,2'-=
diylbis[diphenylphosphine]-*P*,*P*'][(1,2,5,6-=
η)-1,5-cyclooctadeiene]-
chloride, stereoisomer *[101627-26-3]*, 168628a
——, [[1,1'-binaphthalene]-2,2'-diylbis=
[diphenylphosphine]-*P*,*P*'][(1,2,5,6-η)-1,5-=
cyclooctadiene]-
stereoisomer, perchlorate *[82822-45-5]*, 168628a

$C_{52}H_{48}CoN_4O_2S$
Cobalt, (1-butanol)(ethyl mercaptoacetato-*S*)[5,=
10,15,20-tetraphenyl-21*H*,23*H*-porphinato=
(2-)-N^{21},N^{22},N^{23},N^{24}]-
(*OC*-6-23)- *[100203-75-6]*, 64759c

FIGURA 77
Exemplo de entradas do índice de fórmulas do *Chemical Abstracts*
Reproduzido com permissão do Chemical Abstracts Service

Por exemplo, o antigo *British Technology Index (BTI)*, conforme foi descrito no capítulo 4, utilizava entradas de índice formadas por uma cadeia de termos controlados numa 'ordem sistemática'. Veja-se exemplo disso na figura 22 (capítulo 4). Uma referência bibliográfica aparecia somente em um único lugar do índice, o qual era determinado pela seqüência em que os termos eram combinados. Outras possibilidades eram criadas mediante um mecanismo sistemático de remissivas baseado nos princípios da indexação em cadeia. Por exemplo,

88S/037 **Compulsory education and home schooling: truancy or prophecy?** M. A. PITMAN. *Education and Urban Society*, 19(3), 1987, pp 280—289.

Starting from the premise that American schooling is experiencing a crisis of meaning, the author looks at the increased incidence of in-school truancy or class cutting, and the increase in home schooling. Approximately 25 percent of the school population are educated at home, though at least another 9 percent are persistent truants, and up to 20 percent in-school truants. A variety of research is cited throughout the article. Home schoolers are defined as falling into three main categories: religious; progressive; and academic. Religious concerns centre upon the poor quality of public schooling, the moral education of the children and a desire for closer parent-child relationships. The author has carried out a survey of a New Age or Progressive community in the northeastern United States, where the emphasis is on Green politics and alternative lifestyles and approaches. For these people, home schooling makes sense as it allows for unorthodox views and treatment to be provided. The academic home-schoolers are concerned about the academic quality (or lack of it) in public schools. Surveys do show that home-schooled children do perform on average better than public school educated children, though the parents themselves tend to be more highly educated than the population at large. Legally, the laws concerning schooling do not compel education; rather they compel attendance, so home-schoolers tend to receive a disproportionate amount of school superintendent time and activity. In the history of society, the emphasis on compulsory attendance is very recent, and is occurring at the precise time when parents are questioning the quality and nature of public education provided. —*NM*

88S/038 **A comparative study of black dropouts and black high school graduates in an urban public school system.** S. B. WILLIAMS. *Education and Urban Society*, 19(3), 1987, pp 311—319.

A sample of 50 black male and female dropouts from an urban southeast Texan school district in 1985–86 is compared with 50 black male and female graduates from the same school in the same year to ascertain significant differences between them. Data was collected from records, tests and home visits. All the students lived in the attendance zone for the school, which provided an homogeneous socioeconomic background. The researcher was a participant observer, having been a resident in the community for 30 years. Church attendance was found to be a significant factor, with 72 percent of the graduates and 14 percent of the dropouts attending. Graduate status, however, did not help the students in gaining social security assistance. There was a higher incidence of detentions and grade retentions (being kept down a year) for the dropouts than for the graduates, and a lower attendance at vocational educational programmes. Though there was no significant difference in the occupational levels of parents, the parents of the graduates were more highly educated. Similar sibling attainment, and the friendship of other graduates were also significant factors in the background of the graduates. The graduates also had more positive views towards the school than the dropouts, who felt alienated and on the periphery of school and community life. The dropout experiences pervasive feelings of isolation, disconnectedness and rejection, and these must be addressed if the dropout is to be rehabilitated to schooling. —*NM*

FIGURA 78
Exemplo de resumos de *Sociology of Education Abstracts*
Reproduzido com permissão de Taylor & Francis
<http://www.tandf.co.uk>

FIGURA 79
Exemplo de entradas de índice do *Sociology of Education Abstracts*
Reproduzido com permissão de Taylor & Francis <http://www.tandf.co.uk>

usaram-se remissivas do tipo *see* [ver] para gerar pontos de acesso alternativos para os itens sobre 'fabrics' [tecidos] exemplificados na figura 22 (a partir de termos como 'finishing' [acabamento], 'dyeing' [tingimento], 'laminating' [laminação], etc. Note-se também como este índice liga entre si termos considerados semanticamente relacionados ('related headings' [cabeçalhos relacionados]). Embora os princípios em que se baseia a indexação tenham permanecido os mesmos, uma versão posterior dessa publicação, denominada *Current Technology Index* (*CTI*), adotou um método algo diferente de apresentação das referências. Esta modificação foi adotada para economizar espaço e evitar as páginas com uma composição muito sobrecarregada que eram características do *BTI*. As diferenças de leiaute entre o *BTI* e o *CTI* são exemplificadas na figura 81.

Este índice encontra-se hoje em seu terceiro formato, que inclui resumos, e o título atual é *Abstracts in New Technologies and Engineering*. A inclusão de

FIGURA 80
Exemplo de entradas do índice de assuntos do *Epilepsy Abstracts*
Reproduzido com permissão de Elsevier Science Publishers
Este índice é característico dos índices de assuntos produzidos na série *Excerpta Medica*

resumos exigiu uma grande mudança de formato, e a publicação agora se assemelha muito com o formato atual do *Library and Information Science Abstracts*.

Vários índices impressos adotaram o PRECIS (Preserved Context Index System). Um exemplo foi o *British Education Index*. Na figura 82 encontram-se exemplos de entradas dessa publicação. Uma referência bibliográfica aparecia sob todos os termos 'importantes' que ocorressem num enunciado de assuntos, cada um deles sendo 'desviado' [*shunted*] para a posição de entrada conforme descrito no capítulo 4. Por exemplo, a segunda entrada para 'agressão' [aggression], na figura 82, é repetida sob 'Pupils' [alunos] e sob 'Primary schools' [escolas primárias]. Desde 1986, o PRECIS não é mais utilizado como base da indexação do *British Education Index*.

BTI Heading STEEL : Production : Furnaces, Arc : Ladles
References LADLES : Arc furnaces : Steel production.
See STEEL : Production : Furnaces, Arc : Ladles
ARC FURNACES : Steel production. See STEEL : Production : Furnaces, Arc
FURNACES, Arc : Steel production. See STEEL : Production : Furnaces, Arc

CTI Heading STEEL : Production : Furnaces, Arc : Ladles
References LADLES
See
Steel : Production : Furnaces, Arc : Ladles
ARC FURNACES
See
Furnaces, Arc
FURNACES, Arc
See
Steel : Production : Furnaces, Arc

FIGURA 81
Diferenças na apresentação de referências entre o *British Technology Index (BTI)* e o *Current Technology Index (CTI)* de um item sobre cadinhos [*ladles*] para fornos a arco elétrico [*arc furnaces*] na produção de aço [*steel*]
O autor agradece a Tom Edwards, ex-editor do *Current Technology Index*, por este exemplo.
Ambos os exemplos são reproduzidos com a gentil permissão de CSA

AGGRESSION
See also
Violence
AGGRESSION. Children
Coping by adults
Coping with physical violence : some suggestions / John Jamieson. — *Mal. Ther. Educ.*, Vol.2, no.2 : Autumn 84. — p39-45
Bibliography: p45

AGGRESSION. Pupils. Primary schools
Identification
Identification of aggressive behaviour tendencies in junior age children : first stage in a study of aggression / C. Gilmore ... [et al.]. — *Educ. Rev.*, Vol.37, no.1 : Feb 85. — p53-63
Bibliography: p63

AGRICULTURAL COLLEGES
Curriculum. Innovation — Australasia — Case studies
Learning to be a capable systems agriculturist / Richard Bawden and Ian Valentine. — *Program. Learn. Educ. Technol.*, Vol.21, no.4 : Nov 84. — p273-287
Education for Capability. — Bibliography: p286-287

AGRICULTURAL COLLEGES
Management (curriculum subject). Courses. Development — Nigeria
Development of management courses for the agriculture sector in Nigeria / A.E. Shears. — *Program. Learn. Educ. Technol.*, Vol.21, no.2 : May 84. — p88-94
Dissemination and Diffusion. — Bibliography: p94

AGRICULTURAL COLLEGES
Teaching aids: Microcomputer systems — Case studies
Computers in agricultural education / by Andrew Todd. — *Comput. Educ.*, No.48 : Nov 84. — p24

AGRICULTURAL LECTURERS
Lecture notes. Inclusion of new material — Case studies
Sources of new materials included in lectures by lecturers in agriculture / J.T. Smith, B.W. Rockett
Bibliography: p299
Pt 2: An analysis of published sources used. — *High. Educ.*, Vol.13, no.3 : Jun 84. — p289-299

FIGURA 82
Exemplo de entradas PRECIS do *British Education Index*
Reproduzido com permissão da British Library

Índices de citações

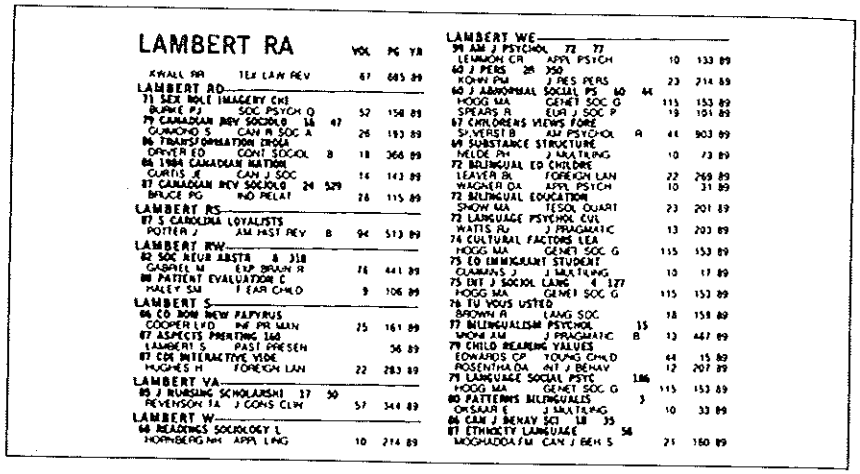
O Institute for Scientific Information (ISI) publica atualmente três índices de citações: o *Science Citation Index*, o *Social Sciences Citation Index* e o *Arts and Humanities Citation Index*. Em virtude de serem bastante diferentes dos outros índices impressos descritos neste capítulo, merecem atenção à parte.

A utilidade fundamental de um índice de citações é encontrar para determinado item bibliográfico, que seja do conhecimento de quem faz a busca, itens posteriores que o citaram. A figura 83 apresenta alguns exemplos de entradas do *Social Sciences Citation Index* (os outros índices de citações obedecem aos mesmos princípios). Suponhamos que sabemos que um artigo de W.E. Lambert, que começa na página 44 do *Journal of Abnormal and Social Psychology*, volume 60, 1960, é altamente relevante para um interesse de pesquisa atual. Buscando no *SSCI* sob o nome do autor (figura 83) localizamos esse artigo e encontramos outros, posteriores a ele, que o citaram. Neste exemplo o artigo é citado por dois outros itens publicados em 1989 (por Hogg e por Spears).

A figura 83 foi extraída da seção *Citation Index* [índice de citações] do *Social Sciences Citation Index*. Observe-se que, sob o nome de cada autor, as entradas aparecem em ordem de data de publicação. Para os itens citantes apresentam-se apenas sucintas informações bibliográficas. Para conseguir dados bibliográficos mais completos devemos nos dirigir a outra seção do *SSCI*, o *Source Index* [índice de fontes]. Por exemplo, o item citante da autoria de Spears foi publicado no *European Journal of Social Psychology*, volume 19, 1989, e começa na página 101. Para obter informações bibliográficas mais completas (título e números de páginas completos) temos de procurar sob seu nome no *Source Index*.

Os índices de fontes do *Social Sciences Citation Index* e do *Arts and Humanities Citation Index* (porém não do *Science Citation Index*) fornecem, para cada item incluído, uma lista das referências bibliográficas que aparecem no final do artigo (ver, por exemplo, a figura 84).

Nos índices de citações, uma forma original de índice de palavras-chave oferece uma abordagem temática dos itens citantes (fontes). Denominado *Permuterm Subject Index* [índice de assuntos Permuterm], baseia-se em palavras-chave que ocorrem nos títulos dos itens citantes. A figura 85 mostra um exemplo de entrada sob termos que começam com a raiz 'debt' [dívida], conforme aparecem nos títulos de diversos itens citantes. Note-se que são empregadas algumas palavras compostas (por exemplo, 'debt-financed' [financiado pela dívida], bem como palavras simples. Cada entrada mostra, em ordem alfabética, outras palavras-chave que tenham ocorrido junto com ela nos títulos dos itens citantes. Assim, um item sob DEBTS (de autoria de Giguere) trata das dívidas do Terceiro Mundo, outro (de autoria de Garfield) trata das dívidas intelectuais, e assim por diante. Observe-se que as entradas se repetirão sob cada palavra-chave importante do título (por exemplo, uma entrada sob a pala-



LAMBERT RA

KWALL RA TEX LAW REV 67 685 89

LAMBERT RD
71 SEX ROLE IMAGERY CHI
BURKE PJ SOC PSYCH Q 52 158 89
79 CANADIAN REV SOCIOLO 16 47
CLIMO S CAN R SOC A 26 193 89
86 TRANSFORMATION ENGLA
DRIVER ED CONT SOCIOL 8 18 366 89
86 1984 CANADIAN NATION
CURTIS J CAN J SOC 14 143 89
87 CANADIAN REV SOCIOLO 24 529
BRUCE PG MID PELAT 28 115 89

LAMBERT RS
87 S CAROLINA LOYALISTS
POTTER J AM HIST REV 8 94 513 89

LAMBERT RW
82 SOC REVR ABSTR 8 318
GABRIEL M EXP BRAIN R 76 441 89
88 PATIENT EVALUATION C
HALEY SM J EAR CHILD 9 106 89

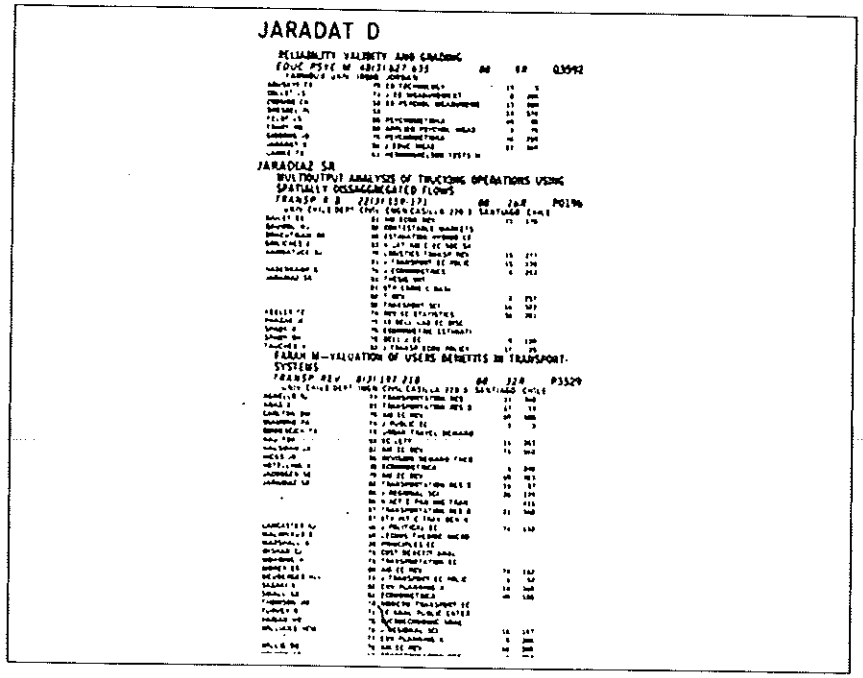
LAMBERT S
86 CD ROM NEW PAPYRUS
COOPER LVD PHI PSI MAN 25 161 89
87 ASPECTS PRINTING 168
LAMBERT S PAST PRESEN 56 89
87 CDE INTERACTIVE VIDE
HUGHES M FOREIGN LAN 22 283 89

LAMBERT VA
85 J NURSING SCHOLARSHI 17 50
REVENSON TA J CONS CLIN 57 344 89

LAMBERT W
66 READINGS SOCIOLOGY L
HORNBERG NH APPL LING 10 214 89

LAMBERT WE
59 AM J PSYCHOL 72 77
LEMMON CR APPL PSYCH 10 133 89
60 J PERS 28 350
KOHN PH J PES PERS 23 214 89
60 J ABNORMAL SOCIAL PS 60 44
HOGG MA GENET SOC G 115 153 89
SPEARS R EUR J SOC P 19 191 89
67 CHILDRENS VIEWS FORE
SEVERS B AM PSYCHOL A 44 903 89
69 SUBSTANCE STRUCTURE
MELDE PH J MULTILING 10 73 89
72 BILINGUAL ED CHILDRE
LEAVER BL FOREIGN LAN 22 269 89
WAGNER DA LEPOL PSYCH 10 31 89
72 BILINGUAL EDUCATION
SHOW MA TESOL QUART 23 201 89
72 LANGUAGE PSYCHOL CUL
WATTS RJ J PRAGMATIC 13 203 89
74 CULTURAL FACTORS LEA
HOGG MA GENET SOC G 115 153 89
75 ED IMMIGRANT STUDENT
CUMMINS J J MULTILING 10 17 89
75 INT J SOCIOL LANG 4 127
HOGG MA GENET SOC G 115 153 89
76 TV YOUS USTED
BROWN R LANG SOC 18 158 89
77 BILINGUALISM PSYCHOL 35
SPOLSKY AM J PRAGMATIC 8 13 467 89
79 CHILD REARING VALUES
EDWARDS CP YOUNG CHILD 44 15 89
ROSENTHAL DA AUT J BEHAV 12 207 89
79 LANGUAGE SOCIAL PSYC 186
HOGG MA GENET SOC G 115 153 89
80 PATTERNS BILINGUALIS 5
OKSAAR E J MULTILING 10 33 89
86 CAN J BEHAV SCI 18 35
87 ETHNICITY LANGUAGE 56
MOGHADDAM CAN J BEH S 21 180 89

FIGURA 83
Exemplo de entradas do *Social Sciences Citation Index*



JARADAT D

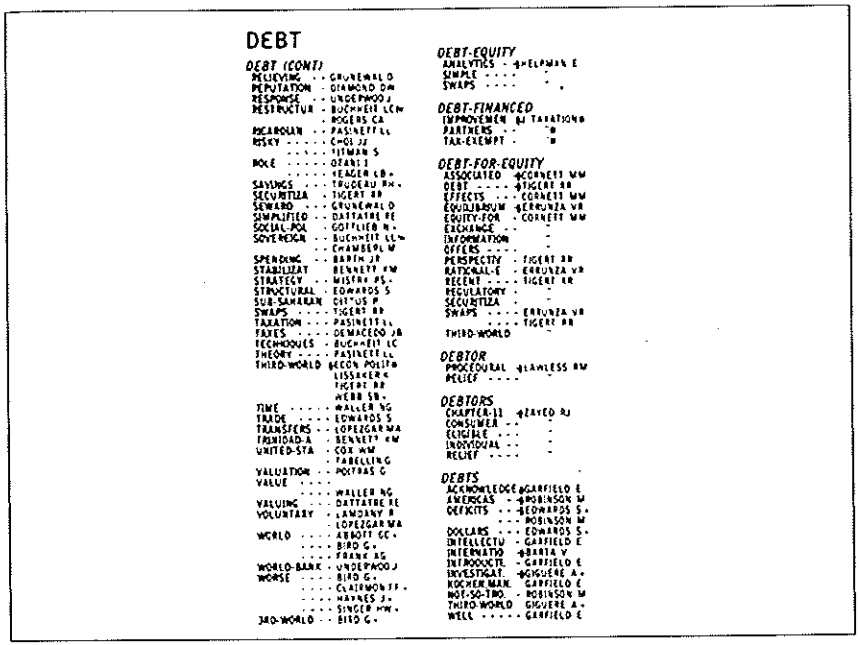
RELIABILITY VALIDITY AND CALIBRA
FOUC PSYC M 48(3):627-635 90 89 03592

JARADAT SA
MULTIOUTPUT ANALYSIS OF TRUCKING OPERATIONS USING SPATIALLY DISAGGREGATED FLOWS
TRANSP R B 23(3):159-171 88 JUN P0196

FARAH M—VALUATION OF USERS BENEFITS IN TRANSPORT-SYSTEMS
TRANSP REV 8(3):197-218 88 JUN P3529

FIGURA 84
Exemplo de entrada do índice de fontes do *Social Sciences Citation Index*

vra-chave 'Third World' [Terceiro Mundo] será modificada pelo termo 'debts' [dívidas]. É evidente que a eficácia deste tipo de índice de assuntos depende inteiramente da qualidade descritiva dos títulos usados na sua geração e da habilidade de quem faz a busca, uma vez que não se adota nenhuma forma de controle de vocabulário.



DEBT

DEBT (CONT)
BELIEVING - - GRUNEWALD
REPUTATION - GRUNWALD DM
RESPONSE - - UNDERWOOD J
RESTRUCTUR - BUCHHEIT LCM
- ROGERS CA
RICARDIAN - - PASINETTI L
RISKY - - - - CHOI JJ
- - - - - TITMAN S
ROLE - - - - - DEANE I
- - - - - YEAGER LB
SAVINGS - - TRUDEAU BM
SECURITIZA - TIGERT RR
SEWARD - - - GRUNEWALD
SIMPLIFIED - DATTATRE RE
SOCIAL-POL - GOTTLIEB N
SOVEREIGN - - BUCHHEIT LCM
- - CHAMBERLN M
SPENDING - - BARTH JR
STABILIZAT - BENNETT RW
STRATEGY - - MISRA PS
STRUCTURAL - EDWARDS S
SUB-SAHARAN CHTUS P
SWAPS - - - TIGERT RR
TAXATION - - PASINETTI L
TAXES - - - DEMARCEGO JB
TECHNIQUES - BUCHHEIT LC
THEORY - - - PASINETTI L
THIRD-WORLD BECON POLITA
LISSAKER K
TIGERT RR
WEBB SR
TIME - - - - - WALLER NG
TRADE - - - - EDWARDS S
TRANSFERS - - LOPEZGARMA
TRINIDAD-A - BENNETT RW
UNITED-STA - COX WW
- TABELLINI G
VALUATION - - POITRAS G
VALUE - - - -
- - - - - WALLER NG
VALUING - - - DATTATRE RE
VOLUNTARY - - LAMDANY R
- LOPEZGARMA
WORLD - - - - ABBOTT GC
- - - - BIRD G
- - - - FRANK AG
WORLD-BANK - UNDERWOOD J
WORSE - - - - BIRD G
- - - - CLAIRMONT F
- - - - HAYNES J
- - - - SINGER HW
3RD-WORLD - - BIRD G

DEBT-EQUITY
ANALYTICS - KELPMAN E
SIMPLE - - - -
SWAPS - - - -

DEBT-FINANCED
IMPROVEMEN - TAXATION
PARTNERS - -
TAX-EXEMPT - -

DEBT-FOR-EQUITY
ASSOCIATED - CORNETT MM
DEBT - - - - TIGERT RR
EFFECTS - - - CORNETT MM
EQUILIBRIUM - ERRUNZA VR
EQUITY-FOR - CORNETT MM
EXCHANGE - -
INFORMATION -
OFFERS - - - -
PERSPECTIV - TIGERT RR
RATIONALE - ERRUNZA VR
RECENT - - - TIGERT RR
REGULATORY -
SECURITIZA -
SWAPS - - - - ERRUNZA VR
- - - - TIGERT RR
THIRD-WORLD -

DEBTOR
PROCEDURAL - SEAMLESS RM
RELIEF - - - -

DEBTORS
CHAPTER-11 - REAVED RJ
CONSUMER - -
ELIGIBLE - - -
INDIVIDUAL - -
RELIEF - - - -

DEBTS
ACKNOWLEDGE - GARFIELD E
AMERICAS - - ROBINSON M
DEFICITS - - - EDWARDS S
- - - ROBINSON M
DOLLARS - - - EDWARDS S
INTELLECTU - GARFIELD E
INTERNATIO - BARRA V
INTRODUCTI - GARFIELD E
INVESTIGAT - GIGUERE A
KOCHER MAN - GARFIELD E
NOT-SO-TRO - ROBINSON M
THIRD-WORLD - GIGUERE A
WELL - - - - - GARFIELD E

FIGURA 85
Exemplo de entrada do índice de assuntos Permuterm do *Social Sciences Citation Index*

As várias partes que compõem esses índices de citações fazem com que sejam poderosas ferramentas de busca bibliográfica. Eles ensejam diferentes métodos de busca. Pode-se iniciar uma busca com a referência bibliográfica de um item sabidamente de interesse ou começá-la com uma palavra-chave. As palavras-chave levam a outras palavras-chave possíveis e os títulos dos itens citantes também sugerem palavras-chave adicionais que seriam úteis na busca. Tomando-se um exemplo hipotético, uma busca por palavra-chave no *SSCI* de 1996 levaria a um item altamente relevante que seria investigado visando à identificação de itens posteriores que o tivessem citado. Estes, por sua vez, poderiam sugerir outras palavras-chave que levariam a outros documentos que também seriam investigados em busca de citações posteriores, e assim sucessivamente numa série de iterações. Nos índices de citações em que o índice de

fontes inclui as referências bibliográficas (ver figura 84), são possíveis outras formas de iteração. Por exemplo, uma busca sobre um item sabidamente de alta relevância pode levar a um item citante altamente relevante. Algumas das referências no item citante serão então investigadas para localizar outros itens que as citem, e assim sucessivamente.

Os índices de citações impressos possuem bases de dados equivalentes em formato eletrônico. Estes e muitos outros índices mencionados neste capítulo, são hoje acessíveis pela Rede. O princípio da citação — um item bibliográfico que cita (referencia) um anterior — também pode ser adotado para ligar publicações por outros meios — mediante acoplamento bibliográfico ou co-citação (ver capítulo 15).

Outro produto bastante conhecido do Institute for Scientific Information é o *Current Contents*, publicação semanal, editado em várias seções que abrangem diferentes assuntos, que reproduz as páginas de sumários de uma ampla gama de periódicos. A figura 86 mostra um exemplo. Cada fascículo do *Current Contents* inclui um índice de palavras-chave bastante simples, como mostra a figura 87; um dos termos desse exemplo (*glucose*) [glicose] tem relação com um dos itens da figura 86. Observe-se que o índice inclui algumas expressões e nomes, bem como palavras-chave simples. Cada entrada leva a uma página do *Current Contents* e a um número de página do periódico ali representado. Por exemplo, uma das entradas sob 'glucose' remete ao item que começa na página 3214 do fascículo de dezembro de 1989 de *Applied and Environmental Microbiology* (figura 86). Este índice simples é usado de duas formas. Evidentemente, pode-se simplesmente investigar todas as referências a determinada palavra-chave. No entanto, um especialista em buscas mais experiente, que estiver procurando informações mais específicas, poderá optar por combinar palavras-chave. Por exemplo, se alguém estivesse buscando artigos sobre glicose no contexto de leveduras, compararia os números que aparecem sob o termo *glucose* [glicose] com os que aparecem sob *yeast* e *yeasts* [levedura, leveduras], para verificar se algum número ocorre sob ambos os termos. Em caso positivo, talvez esse número se refira a itens que tratam precisamente do tópico da busca, inclusive um dos artigos que aparecem na figura 86. Isso corresponde, basicamente, a uma variante do sistema Uniterm (ou pelo menos a implementação desse sistema na prática), conforme se mencionou no capítulo 2. O sistema Uniterm foi uma das primeiras formas de sistema de recuperação pós-coordenado.

Conclusão

Neste capítulo foram exemplificados diferentes métodos de implementação de um serviço de indexação/resumos em formato impresso. Embora umas pessoas prefiram um método e outras pessoas prefiram outro, nenhum método é, *ipso facto*, melhor do que o resto. Isso depende muito de como o serviço será utilizado.

CC666 **Applied and Environmental Microbiology** Articles and Abstracts in English — Amer Soc Microbiol

VOL. 55 NO. 12 DECEMBER 1989 (L,A)

GENETICS AND MOLECULAR BIOLOGY

ENZYMOLOGY AND PROTEIN ENGINEERING

PHYSIOLOGY AND BIOTECHNOLOGY

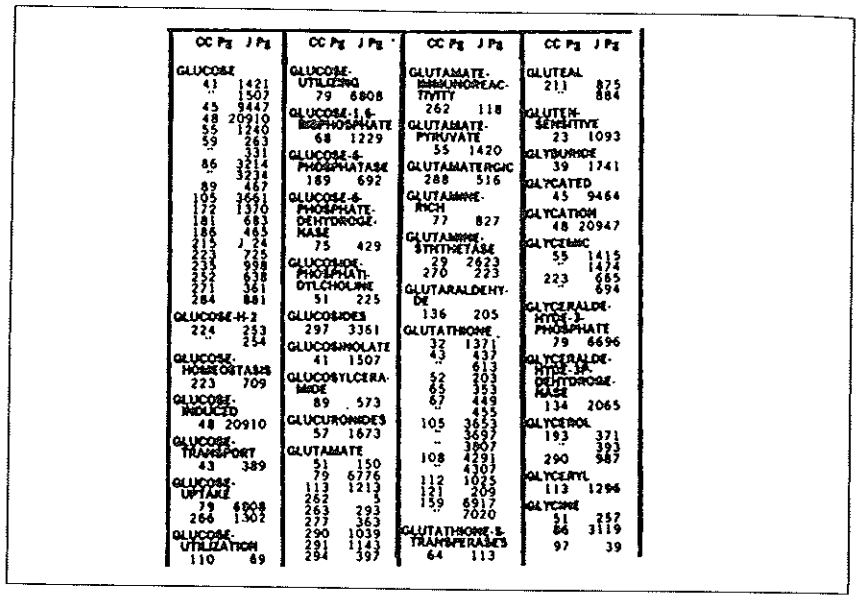
FOOD MICROBIOLOGY

MYCOLOGY

CONTINUED

FIGURA 86

Exemplo de página do *Current Contents*



CC Pg J Pg

GLUCOSE
41 1421
" 1507
45 9447
48 20910
55 1240
59 263
" 331
86 3214
" 3234
89 467
105 3661
172 1370
181 683
186 484
215 J 24
223 725
233 998
252 638
271 361
284 881

GLUCOSE-H-2
224 253
" 254

GLUCOSE-HOMEOSTASIS
223 709

GLUCOSE-INDUCED
48 20910

GLUCOSE-TRANSPORT
43 389

GLUCOSE-UPTAKE
79 6808
266 1302

GLUCOSE-UTILIZATION
110 69

GLUCOSE-UTILIZING
79 6808

GLUCOSE-1,6-BISPHOSPHATE
68 1229

GLUCOSE-6-PHOSPHATASE
189 692

GLUCOSE-6-PHOSPHATE-DEHYDROGENASE
75 429

GLUCOSIDE-PHOSPHATIDYLCHOLINE
51 225

GLUCOSIDES
297 3361

GLUCOSINOLATE
41 1507

GLUCOSYLCERAMIDE
89 573

GLUCURONIDES
57 1673

GLUTAMATE
51 150
79 6776
113 1213
262 3
263 293
277 363
290 1039
291 1143
294 397

GLUTAMATE-IMMUNOREACTIVITY
262 118

GLUTAMATE-PYRUVATE
55 1420

GLUTAMATERGIC
288 516

GLUTAMINE-RICH
77 827

GLUTAMINE-SYNTHETASE
29 2923
270 223

GLUTARALDEHYDE
136 205

GLUTATHIONE
32 1371
43 437
" 613
52 203
65 353
67 449
" 455
105 3653
" 3697
" 3807
108 4291
" 4307
112 1025
121 209
159 6917
" 7020

GLUTATHIONE-S-TRANSFERASES
64 113

GLUTEAL
211 875
" 884

GLUTEN-SENSITIVE
23 1093

GLYBURIDE
39 1741

GLYCATED
45 9464

GLYCATION
48 20947

GLYCEMIC
55 1415
" 1474
223 665
" 694

GLYCERALDEHYDE-3-PHOSPHATE
79 6696

GLYCERALDEHYDE-3P-DEHYDROGENASE
134 2065

GLYCEROL
193 371
" 393
290 987

GLYCERYL
113 1296

GLYCINE
51 257
86 3119
97 39

FIGURA 87
Exemplo de entradas do índice de palavras-chave do *Current Contents*

Para atender às finalidades da notificação corrente [serviço de alerta], as ferramentas que empregam alguma forma de método classificado normalmente serão superiores aos índices alfabético-específicos, pelo menos na medida em que o esquema de classificação corresponda aos interesses de um grupo de usuários. Por exemplo, alguém interessado em se manter a par dos novos avanços no campo da parasitologia em geral certamente achará o *Biological Abstracts*, que dedica uma seção a este tópico, mais útil do que o *Index Medicus*, onde as referências ao assunto provavelmente estarão dispersas sob uma ampla variedade de cabeçalhos de assuntos. No entanto, para alguém que tenha interesse em se manter notificado correntemente sobre assuntos altamente específicos, o método alfabético-específico seria, de fato, mais conveniente. Por exemplo, o *Index Medicus* provavelmente seria um instrumento muito útil para se manter atualizado a respeito da bibliografia sobre retinite pigmentosa, por exemplo.

Ao considerar esses diversos instrumentos como dispositivos de busca e recuperação, é evidente que passam a ter influência nisso todos os fatores de desempenho analisados em outra parte deste livro. Quer dizer, a eficácia de um índice impresso como ferramenta de busca dependerá do número de pontos de acesso que proporcionar, da especificidade do vocabulário empregado na indexação, da qualidade e coerência da indexação e da extensão com que essa ferramenta ofereça ajuda positiva a quem faz a busca (por exemplo, por meio da ligação entre si de termos semanticamente relacionados). Devido ao fato de os índices da *Excerpta Medica* proporcionarem maior número de pontos de acesso temático por item do que o *Index Medicus*, provavelmente propiciarão uma melhor revocação. Por outro lado, como o *Index Medicus* indexa cada item somente sob os termos 'mais importantes', é bem provável que ofereça maior precisão.

Os serviços impressos que incluem resumos são superiores aos que não os incluem, em virtude de proporcionarem mais informações que ajudam o usuário a decidir se determinado item realmente lhe será útil. Isso é especialmente valioso no caso de itens que sejam difíceis de encontrar ou de itens escritos em línguas desconhecidas por parte de quem faz a busca. No entanto, os resumos nem sempre são essenciais. Por exemplo, a combinação do título de um item com o cabeçalho de assunto e o subcabeçalho sob os quais aparece, como no exemplo do *Index Medicus*, freqüentemente basta para indicar sua relevância potencial.

Finalmente, é óbvio que índices baseados apenas nas palavras do título proporcionam um método de recuperação bastante limitado. Contudo, mesmo esses índices têm suas vantagens. Por exemplo, uma busca altamente específica que envolva, digamos, um nome próprio, realmente seria mais fácil de ser efetuada num índice baseado em palavras do título do que num outro baseado num vocabulário controlado de caráter genérico. Além disso, quando se recuperam itens durante buscas baseadas em palavras-chave dos títulos, e desde que a palavra-chave seja altamente específica, existe uma chance muito grande de esses itens serem 'relevantes'.

Em virtude de a maioria dos bibliotecários e outros profissionais da informação ser de opinião que os índices impressos são mais fáceis de usar, muitas vezes eles pressupõem que essa seja uma verdade universal. De fato, inúmeros estudos demonstraram que o público leigo pode enfrentar dificuldades ao usar ou mesmo compreender o 'mais simples' dos índices, como os que vêm no final dos livros (ver, por exemplo, Liddy e Jörgensen, 1993a,b).

Durante a última década, mais ou menos, foi feito um esforço visando a tornar os serviços de indexação e resumos mais 'simples', como se viu pelo abandono de um método de classificação facetada no *Library and Information Science Abstracts* e do PRECIS no *British Education Index*. No entanto, tornar essas ferramentas mais amigáveis para o usuário talvez não seja a salvação delas. O fato de muitas bibliotecas estarem cancelando as assinaturas das edições impressas, dando preferência ao acesso às versões eletrônicas, sugere que fontes desse tipo em formato impresso talvez tenham hoje uma expectativa de vida muito curta.

CAPÍTULO 11

Como melhorar a indexação

Em todo este livro, deixou-se explícito, em geral, que o resultado final da indexação de um documento é uma simples lista de termos, às vezes selecionados de um vocabulário controlado, que, em conjunto, descrevem o conteúdo temático analisado no documento. Com freqüência, todos os termos dessa lista são considerados em pé de igualdade (isto é, o indexador não especifica que alguns são mais importantes do que outros) e, comumente, não se identificam quaisquer relações explícitas entre os termos.

A indexação, porém, é um pouco mais complexa do que isso: aos termos podem ser atribuídos pesos que reflitam a percepção que o indexador tem de sua importância, e/ou ser feito um esforço no sentido de acrescentar um pouco de 'sintaxe' aos termos, de modo que suas inter-relações se tornem mais claras.

Indexação ponderada

Grande parte da indexação de assuntos implica uma simples decisão binária: um termo é ou não é atribuído a um documento. Embora isso simplifique o processo de indexação, cria, efetivamente, alguns problemas para o usuário de uma base de dados, que fica impossibilitado de arquitetar uma estratégia de busca que venha a diferençar itens em que um assunto receba um tratamento substancial daqueles em que o mesmo assunto seja tratado de forma bastante secundária.

Na indexação ponderada, o indexador atribui a um termo um valor numérico que reflete sua opinião sobre a importância desse termo para indicar de que trata determinado documento. Comumente, quanto mais dominante for o assunto, ou mais detalhes o documento contiver a respeito dele, maior será o peso. Vejamos, por exemplo, uma escala numérica de cinco pontos, em que cinco seja o escore mais alto. Aplicando-a ao item exemplificado na figura 3, os termos OPINIÃO PÚBLICA, PESQUISAS POR TELEFONE, ATITUDES e ORIENTE MÉDIO receberiam peso 5, ESTADOS UNIDOS, peso 4, ISRAEL e EGITO, peso 3, e assim por diante. Evidentemente, trata-se de algo subjetivo, e indexadores diferentes atribuirão pesos diferentes. É quase certo, porém, que a maioria atribuiria a ORIENTE MÉDIO um peso alto e a LÍDERES POLÍTICOS ou AJUDA EXTERNA um peso baixo.

Esse tipo de indexação ponderada pode ser adotado de duas formas na recuperação da informação numa base de dados. Uma delas consiste simplesmente em admitir que a pessoa que faz a busca especifique que somente sejam recuperados os itens indexados sob um termo a que foi atribuído determinado peso.

Assim, alguém interessado em artigos que tratem diretamente do assunto 'líderes do Oriente Médio' exigiria que ambos os termos, ORIENTE MÉDIO e LÍDERES POLÍTICOS, contivessem pelo menos um peso igual a quatro. Isso evitaria a recuperação do item exemplificado na figura 3, que trata de líderes políticos de forma bastante secundária, e provavelmente de muitos outros itens como esse.

Uma aplicação alternativa disso é empregar os pesos para ordenar os itens recuperados numa busca. Desse modo, numa busca que exigisse a co-ocorrência de ORIENTE MÉDIO e LÍDERES POLÍTICOS, os itens em que ambos os termos tivessem um peso igual a cinco (peso total igual a dez) seriam impressos ou exibidos em primeiro lugar, vindo em segundo lugar os itens com o escore nove, e assim por diante até os itens cujo escore fosse apenas dois.

Há muito que a atribuição de pesos numéricos aos termos é defendida por Maron (Maron & Kuhns, 1960; Maron et al., 1959; Maron, 1988), que se refere a esse tipo de indexação como 'probabilístico'. Apesar dessa defesa, desconheço qualquer sistema convencional de recuperação (isto é, baseado em indexação feita por seres humanos) que adote pesos numéricos exatamente dessa forma, embora a ponderação de termos esteja implícita em certos sistemas de recuperação automáticos ou semi-automáticos, como o SMART (ver capítulo 15).

No entanto, algumas bases de dados realmente incorporam uma técnica de ponderação simples ao distinguir entre descritores 'mais importantes' e 'menos importantes', o que equivale à adoção de uma escala numérica de dois valores. Esta prática pode ser vinculada à produção de um índice impresso, onde os descritores mais importantes são aqueles sob os quais um item aparece no índice impresso, e os menos importantes são encontrados somente na base de dados em formato eletrônico. É o que fazem, por exemplo, a National Library of Medicine (*Index Medicus* e a base de dados MEDLINE), o National Technical Information Service (NTIS) e o Educational Resources Information Center (ERIC). Mesmo esse método simples de ponderação traz certa flexibilidade às buscas, antes citada. Quem faz as buscas pode especificar que somente sejam recuperados os itens em que um termo (ou termos) apareça como descritor mais importante. Alternativamente, obtém-se uma ordenação incipiente dos resultados, como em:

M * M
M * m
m * m

Isto é, itens em que dois termos, usados por quem faz a busca numa relação *e*, e sendo ambos descritores mais importantes (M), virão em primeiro lugar, seguidos daqueles em que apenas um dos dois seja um descritor mais importante, e depois por aqueles em que ambos sejam somente descritores menos importantes (m).

Alguns serviços de informação ultrapassaram uma escala de ponderação de dois pontos. No BIOSIS, por exemplo, houve época em que eram atribuídos cabeçalhos conceituais [*Concept Headings*] em qualquer um de três 'níveis de ênfase': primário (o item aparece sob este cabeçalho em índices impressos), secun-

dário (ênfase comparativamente forte), e terciário (ênfase secundária) (Vleduts-Stokolov, 1987).

Observe-se que a indexação ponderada, de fato, oferece a quem faz a busca a capacidade de variar a exaustividade da indexação. Voltando à figura 3, é possível que os primeiros cinco termos listados sejam considerados descritores mais importantes, e os nove restantes sejam considerados menos importantes. Nesse caso, a estratégia de busca que especificasse apenas descritores mais importantes equivaleria, com efeito, a fazer a busca em nível de indexação menos exaustivo.

É importante reconhecer a diferença entre *indexação ponderada*, do tipo aqui descrito, e *busca com termos ponderados*. Esta última nada tem a ver com a indexação ponderada. Ao contrário, refere-se à elaboração de uma estratégia de busca cuja lógica é orientada por pesos numéricos e não por operadores booleanos. Por exemplo, a estratégia de busca assumiria o seguinte formato:

Termo	*Peso*	
A	10	
B	10	
C	2	Limiar = 20
D	2	
E	1	
F	1	

O menor peso aceitável é 20, o que significa que os termos *A* e *B* devem estar ambos presentes num registro antes de ser recuperado. No entanto, um registro pode exceder o peso mínimo (limiar) de modo que, compreensivelmente, alguns registros terão um escore de 26 (se todos os seis termos estiverem presentes), outros, 25, e assim por diante. Esses itens com escores elevados viriam em primeiro lugar numa saída impressa. Tem-se assim uma saída ordenada por escores, mesmo sem usar qualquer ponderação dos termos de indexação. Este método de busca em bases de dados foi muito comum em sistemas de processamento em lotes, principalmente nos voltados para a Disseminação Seletiva de Informações (DSI). É, porém, muito menos indicado para buscas no modo em linha.

O método ideal de ponderação implicaria que uma equipe fizesse a indexação (ver capítulo 5) e os termos com que concordassem todos os indexadores teriam peso maior, e os que fossem atribuídos por um indexador teriam peso menor. Villarroel et al. (2002) propõem esse método num ambiente de biblioteca digital. Isso pressupõe um registro de texto completo com um campo destinado a termos atribuídos pelos usuários. Os usuários do registro poriam em destaque partes do texto digital que julgassem importantes e isso levaria à revisão dos pesos relativos aos termos de indexação (ou, de fato, as próprias palavras do texto).

Muitos sistemas 'automáticos' incluem formas de ponderação que permitem a ordenação da saída segundo um critério. Sistemas desse tipo são examinados no capítulo 15. Na maioria dos casos, os sistemas de processamento automático ponderam segundo critérios de freqüência: freqüência de ocorrência de um termo num texto e/ou de ocorrência numa base de dados como um todo; ou outros

métodos que foram experimentados, inclusive o emprego de critérios posicionais (por exemplo, a qual distância um do outro se encontram dois termos num texto). Keen (1991) comparou diferentes métodos e concluiu que a associação de métodos combinados provavelmente ofereça melhores resultados.

Elos entre termos

Ao examinar de novo a figura 3, verificar-se-á que o documento ali representado seria recuperado durante várias buscas para as quais ele não constituiria realmente uma resposta apropriada. Algumas dessas recuperações poderiam ser evitadas com o emprego da indexação ponderada ou com a redução da exaustividade da indexação. Por exemplo, qualquer uma das duas soluções evitaria a recuperação desse documento numa busca de informações sobre líderes políticos em geral, para a qual esse item somente teria uma utilidade muito secundária.

Outras recuperações indesejáveis seriam causadas por *falsas associações*, casos em que os termos que fazem com que um item seja recuperado não têm realmente relação alguma entre si no documento. Um exemplo seria a combinação ESTADOS UNIDOS e LÍDERES POLÍTICOS. É óbvio que o documento não trata de líderes políticos dos Estados Unidos, embora provavelmente seja recuperado numa busca sobre este assunto. Como foi salientado antes, a probabilidade de ocorrência de falsas associações desse tipo aumenta com a extensão do registro (isto é, com o número de pontos de acesso ou com a exaustividade da indexação).

Um meio de evitar falsas associações é estabelecer elos entre os termos de indexação. Quer dizer, o documento é, em certo sentido, segmentado em diversos subdocumentos, cada um deles referindo-se a um assunto separado ainda que possivelmente os assuntos de cada um estejam intimamente relacionados entre si. O documento exemplificado na figura 3 seria subdividido da seguinte forma:

Oriente Médio, Nações Árabes, Líderes Políticos, Israel, Egito, Organização para a Libertação da Palestina
Opinião Pública, Pesquisas por Telefone, Estados Unidos, Atitudes, Oriente Médio
Estados Unidos, Ajuda Externa, Egito, Israel
Conferências de Paz, Oriente Médio, Organização para a Libertação da Palestina

e assim por diante.

Observe-se que todos os termos de cada seqüência guardam relação direta entre si e que alguns aparecem em várias dessas seqüências. Cada uma dessas seqüências — ou elos — é identificada com um caractere alfanumérico incluído na própria base de dados. Num sistema de recuperação em linha isso estaria associado ao número do documento no arquivo invertido. Assim, o documento 12024 pode ser segmentado em 12024/1, 12024/2, 12024/3, e assim por diante. Isso proporciona a quem faz a busca a oportunidade de especificar que dois termos co-ocorram não só no registro do documento mas também em determinado elo dentro desse registro, evitando, portanto, muitas das falsas associações do tipo ESTADOS UNIDOS/LÍDERES POLÍTICOS.

Um tipo especial de segmentação é aplicado a documentos com texto com-

pleto, para reduzir a ocorrência de relações indesejáveis e melhorar a recuperação. Williams (1998) refere-se a isso como 'indexação por trechos' [*passage-level indexing*]. Isso será examinado no capítulo 14.

Indicadores de função

Embora os elos sejam eficazes ao evitar certas recuperações indesejáveis, não resolverão todos os problemas. Alguns termos podem estar diretamente relacionados entre si num documento, e assim aparecerem no mesmo elo, mas não estarem relacionados da forma como quem faz a busca gostaria que estivessem. A figura 3 nos mostra de novo excelente exemplo disso: o item em questão poderia muito bem ser recuperado numa busca sobre atitudes do Oriente Médio em relação aos Estados Unidos, apesar de tratar exatamente da relação oposta.

Para evitar esse tipo de problema (uma *relação incorreta entre termos*) é preciso introduzir certa sintaxe na indexação, a fim de eliminar a ambigüidade. O método 'tradicional' consiste em empregar *indicadores de função* (ou *indicadores relacionais*) — códigos que tornam explícitas as relações entre os termos. A fim de eliminar a ambigüidade do caso Estados Unidos/Atitudes/Oriente Médio, só se necessitaria de dois indicadores de função, os quais seriam indicadores direcionais. Por exemplo, empregar-se-ia a letra *A* para designar a idéia de 'destinatário, alvo ou paciente', e *B* para representar 'emissor, doador, origem'. Neste caso, associar-se-ia a função *A* a *Oriente Médio* e *B* a *Estados Unidos*, uma vez que o primeiro é o alvo das atitudes enquanto o segundo é a origem delas.

Evidentemente, nem todos os problemas de ambigüidade são solucionados com o emprego de somente duas funções. Se esses problemas se mantiverem, no entanto, num nível relativamente elementar, um número razoavelmente pequeno de indicadores de função resolverá a maioria deles.

Os elos e funções foram introduzidos em sistemas de recuperação, simultaneamente, no início da década de 1960, quando os sistemas pós-coordenados ainda eram relativamente novos e a recuperação informatizada engatinhava. Durante certo período, esteve muito em voga indexar com o emprego tanto de elos quanto funções, em grande parte devido à influência do Engineers Joint Council (EJC), que introduziu um conjunto de indicadores de função (ver figura 88) que teve ampla aceitação. Esse tipo de indexação altamente estruturada não gozou de estima por muito tempo. Não só era muito caro, porque os indexadores precisavam de muito mais tempo para executá-lo, como também ficou evidente ser extremamente difícil de aplicar, com coerência, os indicadores de função. Se já é muito difícil (ver capítulo 5) obter coerência com métodos de indexação relativamente simples, essa dificuldade aumenta enormemente quanto mais explícito o indexador tiver de ser ao expressar as relações entre os termos. Os problemas não são tão grandes quando se raciocina somente com dois ou três termos ao mesmo tempo. Amiúde, porém, é muito difícil identificar todas as relações aplicáveis a um grupo maior de termos. Ademais, o acréscimo de um termo a um

8	8
O tópico fundamental em exame é; o assunto principal em estudo é; o assunto relatado é; o principal tópico em discussão é; encontra-se uma descrição de	
1	1
Insumo; matéria-prima; material de construção; reagente; metal de base (para ligas); componentes a serem combinados; constituintes a serem combinados; ingredientes a serem combinados; material a ser perfilado; material a ser moldado; minério a ser refinado; subconjuntos a serem montados; insumo de energia (somente numa conversão de energia); dados e tipos de dados (somente quando insumos em processamentos matemáticos); um material que está sendo corroído	
2	2
Saída; produto, subproduto, co-produto; resultado, resultante; produtos intermediários; liga produzida; material resultante; mistura ou formulação resultante; material fabricado; mistura fabricada; dispositivo moldado ou formado; metal ou substância refinada; dispositivo, equipamento ou aparelho feito, montado, construído, fabricado, arquitetado, criado; produção de energia (somente numa conversão de energia); dados e tipos de dados (somente como resultados de processamento matemático)	
3	3
Componente indesejável; resíduo; escória; rejeitos (dispositivos fabricados); contaminante; impureza; poluente, adulterante ou tóxico em insumos, ambientes e materiais que passivamente recebem as ações; material indesejável presente; material desnecessário presente; produto indesejável, subproduto, co-produto	
4	4
Usos ou aplicações indicados, possíveis, pretendidos, presentes ou posteriores. A utilidade ou aplicação que o termo teve, tem agora ou terá no futuro. Para ser usado como, em, para ou com; para uso como, em, para ou com; usado como, em, para ou com; para uso futuro como, em, para ou com	
5	5
Ambiente; meio; atmosfera; solvente; portador (material); apoio (num processo ou operação); veículo (material); hospedeiro; absorvente, adsorvente	
6	6
Causa; variável independente ou controlada; fator que influencia; 'X' como um fator que afeta ou influencia 'Y'; o 'X' em 'Y é uma função de X'	
7	7
Efeito; variável dependente; fator influenciado; 'Y' como um fator afetado ou influenciado por 'X'; o 'Y' em 'Y é uma função de X'	
9	9
Recebendo passivamente uma operação ou processo com nenhuma alteração de identidade, composição, configuração, estrutura molecular, estado físico ou forma física; posse como quando precedida pelas preposições de ou em significando posse; localização como quando precedida pelas preposições em, para ou de significando localização; empregado com meses e anos quando localizam informação (não dados bibliográficos) num contínuo de tempo	
10	10
Meios de realizar o tópico de estudo principal ou outro objetivo	
0	0
Dados bibliográficos, nomes próprios de autores, autores e fontes coletivos, tipos de documentos, datas de publicação, títulos de periódicos e outras publicações, outros dados identificadores de fontes, e adjetivos	

FIGURA 88

O sistema de indicadores de função do EJC

Reproduzido com permissão da American Association of Engineering Societies

grupo pode alterar de algum modo as relações, criando a necessidade de mudança nos indicadores de função ou, no mínimo, aumento do número de funções aplicáveis a cada termo. No caso dos indicadores de função do EJC, os problemas se agravavam porque um deles, a função 8, não era absolutamente um indicador relacional, mas, ao contrário, um meio de ponderar o termo mais importante. As pessoas incumbidas das buscas defrontavam tantas dificuldades ao identificar as funções que o indexador teria atribuído a um termo que acabavam, com freqüência, por omitir totalmente as funções, o que equivale a exigir que um termo apa-

reça em qualquer função e nega por completo a utilidade do recurso. Os problemas acarretados pelo emprego de elos e indicadores de função em sistemas de recuperação foram estudados minuciosamente em outros trabalhos (Lancaster, 1964; Sinnett, 1964; Montague, 1965; Van Oot et al., 1966; Mullison et al., 1969).

Ainda mais elaborado do que o método de indexação do EJC, que emprega elos e funções, era o método de 'código semântico' na recuperação introduzido pelo Center for Documentation and Communication Research da Western Reserve University (Perry e Kent, 1958; Vickery, 1959). O código semântico foi aplicado a um sistema de recuperação informatizado, na área de metais, projetado e operado pela Western Reserve para a American Society for Metals.

O sucedâneo do documento era um 'resumo telegráfico'. Este era redigido segundo um formato padronizado, obedecendo a um conjunto de regras, para eliminar variações e complexidades da estrutura fraseológica do inglês. Foram feitos formulários especiais para análise de assuntos, para ajudar o indexador no registro de aspectos importantes do conteúdo temático na forma de resumo telegráfico. Nele, os termos eram codificados mediante um 'dicionário de códigos semânticos'. A base do código semântico era um 'radical' semântico. Os radicais (havia cerca de 250 no sistema) representavam conceitos relativamente genéricos. Cada radical recebia um código de quatro dígitos formado por três caracteres com um espaço para interpolação de um quarto caractere, como nestes exemplos:

C-TL Catalyst [Catalisador]
C-TR Container [Recipiente]
C-TT Cutting and drilling [Corte e perfuração]
D-DD Damage [Dano]
D-FL Deflection [Desvio]

Os termos particulares eram formados pela inserção do 'infixo' de uma letra no radical semântico e talvez o acréscimo de um sufixo numérico. Por exemplo, DADD representava tanto '*wound*' [lesão] quanto '*decay*' [deterioração], onde *D-DD* é o radical semântico de '*damage*' [dano] e o infixo *A* simplesmente representa '*is a*' [é um]. Em outras palavras, 'lesão' é um tipo de dano. Acrescenta-se um sufixo numérico apenas para distinguir termos que possuam radicais e estrutura de infixos idênticos; o sufixo não tem em si mesmo importância semântica.

Na figura 89 está a lista completa de infixos. O uso deles com um radical permite expressar vários matizes de significado. Por exemplo, '*bag*' [saco] e '*barrel*' [barril] eram ambos representados por *CATR*, onde o infixo *A* indica que são *tipos* de recipientes. '*Side wall*' [parede lateral] era representado por *CITR*, onde o infixo *I* indica *parte* de recipiente. Um conceito complexo específico é formado a partir de vários 'fatores semânticos'. Por exemplo, o assunto 'telefone' é expresso por

DWCM.LQCT.MACH.TURN.001

onde

D-CM representa Informação
L-CT representa Eletricidade

M-CH representa Dispositivo
T-RN representa Transmissão

e 001 é o sufixo exclusivo que distingue o termo de outros (por exemplo, o telégrafo) que tenham os mesmos fatores semânticos. Pode-se combinar até quatro códigos semânticos para formar o código de um conceito específico.

A	é um
E	é feito de
I	é parte de
O	é feito de vários
Q	faz uso de, é produzido, por meio de
U	é usado para, produz (amiúde usado [em inglês] para verbos terminados em *ing*)
V	age sobre
W	causa, influenciado por, sofre a ação de (freqüentemente usado [na língua inglesa] para verbos que terminam em *ed*)
X	caracteriza-se pela ausência de
Y	está ligado a, caracterizado por, caracteristicamente
Z	assemelha-se a, mas não é
P	caracteriza-se por um aumento de
M	caracteriza-se por uma redução de

FIGURA 89
Infixos semânticos do sistema da Western Reserve University
Fonte: Aitchison e Cleverdon (1963)

Os termos num resumo telegráfico são relacionados sintaticamente entre si por meio de indicadores de função. Na figura 90 apresenta-se uma lista deles. Um exemplo da aplicação de funções é:

KOV.KEJ	cristal
,KOV.KEJ.KUJ.	metal
,KOV.KEJ.KUJ.	liga
,KOV.KEJ.KUJ.	berílio
,KWV	hexagonal muito denso
,KWV	elástico

que indica que cristais de ligas metálicas, especificamente o berílio, estão de algum modo sendo processados, e suas propriedades são 'hexagonais muito densos' e 'elásticos'. Note-se o emprego, neste sistema, de 'funções companheiras'. KOV e KWV são funções companheiras ou emparelhadas. Se uma é atribuída a um termo, é quase certo encontrarmos sua companheira atribuída a um segundo termo, para ligá-los e indicar a exata relação entre eles. Assim, indica-se que 'cristal', segundo a função KOV, tem uma propriedade que lhe foi atribuída. Essas propriedades atribuídas são 'elástico' e 'hexagonal muito denso', conforme indicado pela função KWV.

Além dos indicadores de função, o sistema adotava um método altamente elaborado de ligação dos termos (e funções) nos resumos telegráficos. Essa ligação era obtida por meio de vários níveis de 'pontuação':

1. *Sublocução*. Termo ao qual se anexava um ou mais indicadores de função.
2. *Locução*. Conjunto de termos proximamente relacionados em determinada relação. Admite-se um número finito de padrões de locução. Por exemplo:

KAM	(processo)
KQJ	(meio de processo)
KAH	(condição de processo)

3. *Frase*. É composta de locuções e também formada segundo esquemas padronizados. Por exemplo, uma frase pode abranger um produto e sua fabricação ou um material testado e as propriedades determinadas para ele.
4. *Parágrafo*. Trata-se de um conjunto de frases e pode ser coextensivo com o próprio resumo. É também usado para distinguir completamente tópicos diferentes num único resumo telegráfico. A figura 91 mostra um resumo telegráfico completo como seria registrado em meio eletrônico, apresentando pontuação, funções e fatores semânticos.

Ao fazer uma busca nesse sistema, a formulação do pedido era convertida numa estratégia composta de fatores semânticos e indicadores de função. Vários 'níveis', correspondentes à pontuação dos resumos telegráficos, eram utilizados para limitar os critérios a termos que ocorressem em certas unidades. Por exemplo, o nível de busca 4 solicita simplesmente que determinado termo esteja associado a determinado indicador de função. Isso corresponde à sublocução na pontuação do resumo telegráfico.

KEJ	material processado
KUJ	componente principal
KIJ	componente secundário
KOV	propriedade atribuída a
KWV	propriedade atribuída
KAM	processo
KQJ	meio de processo
KAH	condição de processo
KUP	propriedade influenciada ou determinada por processo
KAP	propriedade influenciada por KAL
KAL	fator que influencia KAP
KWJ	produto

FIGURA 90
Indicadores de função do sistema da Western Reserve University utilizados na indexação da literatura de metalurgia
Fonte: Aitchison e Cleverdon (1963)

O sistema Western Reserve era bastante engenhoso e expressava matizes de significado muito sutis. Possuía grande flexibilidade. Podiam-se fazer buscas com grande precisão, usando pontuação, funções e fatores semânticos específicos. Alternativamente, permitia buscas com relativa amplitude (para obter alta revocação) ao se ignorar esses dispositivos e usar a estrutura dos códigos semânticos

como recurso de generalização (por exemplo, usando o conceito geral *D-DD* para 'dano' sempre que ocorresse como componente num código complexo).

```
KOV.KEJ.CARS.009.,KOV.KEJ.CARS.006.,KUJ.KEJ.KOV.MATL.
4.□BQE.,-KAM.CUNG.MWTL.PASS.RQHT.003.,KAM.MAPR.
032.,KAH.DACT.001.*,KAH.LAMN.037.,KAH.DACT.001.*,KAH.
LAMN.024.,KAH.DYFL.6X.PAPR.002.*,KAH.PAPR.PYSH.2X.
001.,-KUP.RANG.009.*,KUP.RAPR.225.,KUP.DASM.006.*,KUP.
PYPR.004.,KUP.DYFL.MATN.002.*,KUP.PYPR.004.,KUP.KAP.
PAPR.017.,KUP.KAP.PAPR.010.,KAL.PAPR.004.,KAL.RANG.
009.*,KAL.MAPR.041.,KUP.PAPR.45X.PWSH.2X.TYRM.001.
,KUP.KAP.PAPR.001.*,KUP.KAP.PAPR.PYSH.2X.001.,KUP.
KAL.MAPR.114.,KUP.KAL.MAPR.087.*,KUP.KAL.MAPR.041.
,KUP.KAL.RANG.009.*,KUP.KAL.MAPR.041.,KUP.KAP.MAPR.
032.*,KUP.KAP.PAPR.PYSH.2X.001.,KAL.DYFL.6X.PAPR.002.
*,KAL.PAPR.PYSH.2X.001.,KAL.RANG.009.*,KAL.PAPR.058.
*,KAL.BYSS.3X.RAPR.002.
```

FIGURA 91
Resumo telegráfico armazenado em formato eletrônico
Fonte: Perry e Kent (1958), *Tools for machine literature searching*. Copyright © 1958, John Wiley & Sons, Inc. Reproduzido com permissão de John Wiley & Sons Inc.

Infelizmente, o sistema era excessivamente artificioso para a finalidade a que se destinava. Era de aplicação complicada, e tanto a indexação quanto a formulação da busca eram operações demoradas e dispendiosas. A experiência posterior nos ensinou que, na maioria das aplicações visando à recuperação da informação, não se precisa do nível de complexidade inerente ao sistema Western Reserve. Era um sistema muito complexo e caro para que fosse economicamente viável, e acabou sendo posto de lado pela American Society for Metals em favor de um método mais simples e com melhor relação custo–eficácia.

Subcabeçalhos

O método de indexação altamente estruturado, exemplificado pelo emprego de elos e funções ou pelo código semântico, predominou no início da década de 1960, quando os sistemas informatizados ainda se achavam num estádio de desenvolvimento muito preliminar. Considerava-se imprescindível, então, obter resultados muito precisos na recuperação, evitando-se a qualquer custo recuperar itens irrelevantes. O exemplo absurdo que se colocava com freqüência era o da necessidade de se distinguir entre *Venetian blinds* [janelas venezianas] e *blind Venetians* [venezianos cegos]! O absurdo do exemplo é óbvio: qual a probabilidade de artigos sobre ambos os assuntos aparecerem na mesma base de dados e quanta bibliografia, seja qual for, existe a respeito de venezianos cegos? Hoje em dia, reconhece-se e se aceita o fato de que ocorrerão recuperações indesejáveis, devidas a associações falsas ou espúrias. No entanto, sua ocorrência é comumente tida como se mantendo dentro de limites aceitáveis. Na avaliação do MEDLARS (Lancaster, 1968a), cerca de 18% de aproximadamente 3 000 falhas de precisão que ocorreram em 302 buscas foram causados por relações ambíguas entre termos. Admite-se, comumente, que é melhor aceitar algumas falhas

desse tipo do que tentar evitá-las com o emprego de métodos de indexação mais elaborados e custosos.

Os problemas decorrentes das associações falsas ou ambíguas são atualmente menos graves do que o eram há 30 ou 40 anos porque existe, na maioria dos sistemas, um alto nível de pré-coordenação. Tais problemas são mais comuns em sistemas baseados na indexação com uma única palavra (Uniterm) ou em sistemas baseados na linguagem natural (ver capítulo 14). Como os tesauros incorporaram um nível mais alto de pré-coordenação, diminui a probabilidade de associações falsas ou ambíguas. Tomemos um exemplo simples. Os termos COMPUTADORES e PROJETO, aplicados a um documento, são ambíguos: os computadores estão sendo projetados ou estão sendo aplicados ao projeto de algo diferente? Por outro lado, a combinação mais pré-coordenada

COMPUTADORES
PROJETO DE AERONAVES

é muito menos ambígua, e a combinação

PROJETO DE AERONAVES
PROJETO ASSISTIDO POR COMPUTADOR

parece totalmente inequívoca.

Uma forma de obter alguma pré-coordenação, sem aumentar grandemente o tamanho do vocabulário controlado, é com o emprego de subcabeçalhos. Num sistema pós-coordenado, aplicam-se os subcabeçalhos de forma muito parecida com o modo como são aplicados nos tradicionais catálogos de assuntos das bibliotecas. Os melhores candidatos a subcabeçalhos são aqueles termos que seriam potencialmente aplicáveis a muitos dos outros termos do vocabulário. Assim, um vocabulário de 5 000 descritores, mais 20 subcabeçalhos, gera, teoricamente, 100 000 (5 000 × 20) termos exclusivos. Na prática, porém, cada subcabeçalho talvez seja aplicável somente a determinada categoria de termo, por isso o número de combinações possíveis não seria tão elevado.

Voltando ao exemplo anterior, PROJETO seria um bom candidato a subcabeçalho em certas bases de dados. Assim, COMPUTADORES/PROJETO é bem menos ambíguo do que a combinação PROJETO e COMPUTADORES. Evidentemente, acrescentar um subcabeçalho a um cabeçalho principal (descritor) é uma forma de ligação (elo) muito simples. Com efeito, porém, os subcabeçalhos funcionam praticamente como elos e funções simples ao mesmo tempo. Vejamos a combinação:

AERONAVES/PROJETO
COMPUTADORES

O termo PROJETO não só se acha ligado explicitamente a AERONAVES mas seu emprego como subcabeçalho implica realmente a relação mais provável entre o termo AERONAVES e o termo COMPUTADORES (isto é, que os computadores são empregados como ferramentas de trabalho no projeto de aeronaves).

A National Library of Medicine foi muito bem-sucedida ao empregar subcabeçalhos exatamente dessa forma. Em alguns casos, os subcabeçalhos se complementam entre si. Assim, a combinação

DISEASE X/CHEMICALLY INDUCED [Doença X/Induzida quimicamente]
DRUG Y/ADVERSE EFFECTS [Droga Y/Efeitos adversos]

implica que a doença X foi causada pelo medicamento Y, enquanto a combinação

DISEASE X/DRUG THERAPY [Doença X/Quimioterapia]
DRUG Y/THERAPEUTIC USE [Droga Y/Uso terapêutico]

expressa uma relação completamente diferente entre X e Y.

Embora a principal justificativa para uso de subcabeçalhos dessa forma fosse facilitar a utilização do *Index Medicus* impresso, comprovou-se que eles foram eficazes ao reduzir as ambigüidades também nas buscas na base de dados eletrônica. Ainda que a indexação com combinações de cabeçalhos principais/subcabeçalhos seja indiscutivelmente menos coerente do que a indexação que emprega somente cabeçalhos principais (Lancaster, 1968a), os subcabeçalhos apresentam menos problemas do que os indicadores de função, e, ao contrário destes, são de compreensão imediata por parte dos usuários.

Dispositivos da linguagem de indexação

Esses dispositivos — ponderação, elos e indicadores de função — são considerados *dispositivos de precisão* porque possibilitam que se aumente a precisão durante uma busca numa base de dados. Outros dispositivos, como o controle de sinônimos, por outro lado, são denominados *dispositivos de revocação* porque tendem a melhorar a revocação. A série completa desses dispositivos é às vezes denominada *dispositivos da linguagem de indexação* (Raitt, 1980; Lancaster, 1986). Isso é um pouco enganoso: alguns desses dispositivos, como os subcabeçalhos e o controle de sinônimos, constituem, de fato, componentes essenciais de uma linguagem de indexação, enquanto outros, como os elos ou a ponderação, são bastante independentes da linguagem de indexação. Ou seja, são operações que se aplicam aos termos quando da indexação e não componentes de um vocabulário controlado. Poder-se-ia, com efeito, separar os *dispositivos da linguagem de indexação* dos *dispositivos de indexação*, mas isso seria considerado uma bizantinice.

Os dispositivos de indexação examinados neste capítulo são todos eles dispositivos de precisão, com exceção de certos componentes do código semântico. Fundamentalmente, um dispositivo de precisão aumenta o tamanho do vocabulário empregado na indexação, enquanto um dispositivo de revocação reduz seu tamanho. Por exemplo, uma escala de ponderação de cinco pontos praticamente aumenta o tamanho do vocabulário por um fator de cinco. Ao invés de se ter um único termo, LÍDERES POLÍTICOS, por exemplo, agora se têm cinco termos — LÍDERES POLÍTICOS 5, LÍDERES POLÍTICOS 4, e assim por diante. Os elos e os indicadores de função causam efeito similar.

Outra maneira de examinar isso é em termos do tamanho da classe: os dispositivos de precisão criam um maior número de classes menores, enquanto os dispositivos de revocação criam um número menor de classes maiores (figura 92).

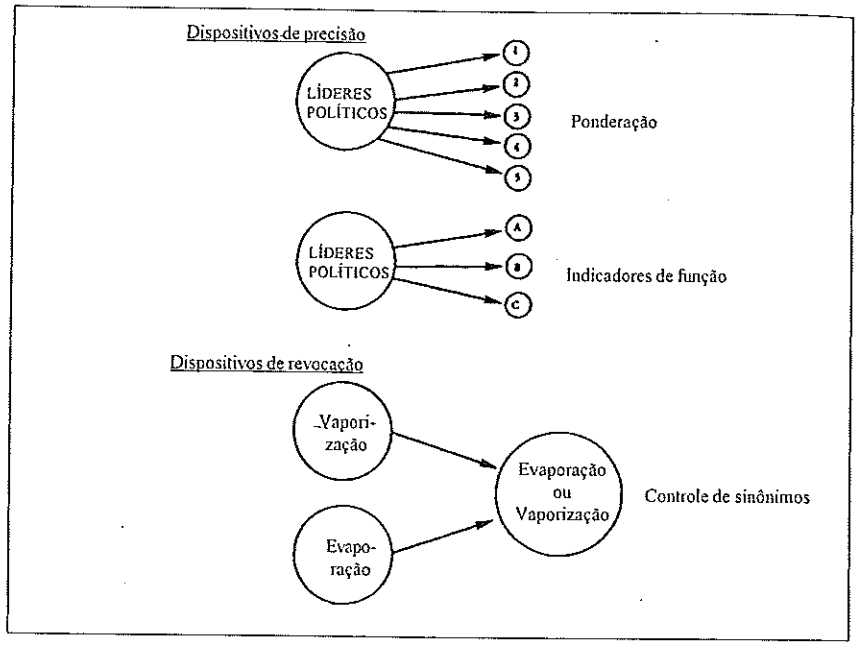


FIGURA 92
Os dispositivos de precisão criam classes menores; os dispositivos de revocação criam classes maiores

Parece provável que a necessidade de uma abordagem altamente estruturada da indexação, especialmente o emprego de alguma forma de indicador relacional, variará de um campo para outro. Isso realmente jamais chegou a ser estudado, embora Green (1997) haja examinado a aplicabilidade de estruturas relacionais à indexação no campo das humanidades.

Existem analogias no processamento informatizado de textos para recuperação (ver capítulos 14 e 15), em que a análise [*parsing*] do texto para evidenciar as subordinações sintáticas equivale ao uso de indicadores de função ou outros indicadores relacionais. Embora essa análise sintática seja provavelmente necessária aos sistemas que procuram responder perguntas a partir diretamente do texto, não existe comprovação real de que ela seja necessária no caso dos requisitos menos rigorosos relativos à recuperação de textos ou passagens de textos. Ademais, a análise sintática por computador ainda está longe de perfeita (McDonald, 1992) e esse nível de processamento seria difícil de justificar, com argumentos de custo–eficácia, na maioria das aplicações de recuperação.

CAPÍTULO 12

Da indexação e redação de resumos de obras de ficção

A indexação de assuntos conta com uma história muito longa, acumulou uma vasta experiência e a bibliografia hoje existente sobre o tema é respeitável. Uma de suas aplicações importantes, no entanto, tem sido bastante negligenciada, pelo menos até muito recentemente: a indexação de trabalhos de ficção, como, por exemplo, obras literárias e filmes. O objetivo deste capítulo é examinar em que medida a indexação e a redação de resumos de uma obra de ficção apresentam problemas similares àqueles implícitos no tratamento de obras 'sérias', tais como manuais, artigos de periódicos ou filmes documentários.

Conforme vimos antes neste livro, a indexação de assuntos envolve duas etapas — 'análise conceitual' e 'tradução' — que são processos intelectuais bastante distintos, ainda que aparentemente realizados como se fossem uma única operação. A etapa de análise conceitual determina 'de que trata' um documento. A questão relativa à 'atinência', no que concerne à indexação, foi examinada alhures, por exemplo, Maron (1977), Hutchins (1978) e Swift et al. (1978), enquanto Pejtersen (1979) e Beghtol (1994) abordaram especificamente a 'atinência' da ficção. Vários aspectos da atinência de textos em geral são analisados por Eco (1979) e Troitskii (1979, 1981).

Esses autores levantam várias questões teóricas ou filosóficas sobre o significado da atinência, as quais não procurarei repetir aqui. Para os objetivos do momento, empregarei a expressão 'trata de' como equivalente a 'capaz de informar sobre'. Ou seja, se certas pessoas puderem aprender algo sobre agricultura por intermédio da leitura de um livro ou assistindo a um filme, eu diria que o livro (filme) 'trata de' agricultura.

A indexação de um filme documentário que aborde alguma técnica agrícola não é essencialmente diferente da indexação de um livro, artigo de periódico ou relatório técnico sobre agricultura. Segundo nossa definição, tudo pode ser considerado como se tratasse de agricultura. Pode, porém, um filme de ficção que por acaso tenha como locação uma fazenda ser considerado como se tratasse de agricultura, principalmente se a agricultura for algo completamente acidental em relação ao enredo do filme? Pode um filme que, de passagem, focaliza a agricultura ser considerado como se tratasse de agricultura? Se, por acaso, o herói de um filme é um agricultor, isto faz com que o filme trate de agricultura? Será que isso o faz ser um filme que trata de agricultores?

A indexação de qualquer tipo de obra de ficção — seja ela uma peça teatral, um romance ou um filme — apresenta problemas que são, realmente, um tanto diferentes dos problemas que envolvem a indexação de obras não-literárias. Os dois tipos são criados com objetivos diferentes: o primeiro, fundamentalmente, para entreter ou suscitar emoções, o segundo, fundamentalmente, para veicular informações. O fato de o primeiro tipo transmitir alguma informação concreta é algo acidental em face do objetivo principal do veículo de comunicação. O fato de o segundo tipo poder, de vez em quando, entreter é igualmente algo fortuito em relação ao objetivo principal do veículo de comunicação.

Se atribuirmos o termo AGRICULTURA a um filme documentário ou a um artigo de periódico sobre este assunto, implicamos que estes itens veiculam certas informações sobre agricultura, e que os usuários de um índice procurarão itens por intermédio deste termo porque querem aprender algo sobre este assunto específico.

Por outro lado, se alguém procura, sob o termo AGRICULTURA, num índice de obras de ficção, por exemplo, filmes, com certeza não o faz com o propósito de encontrar informações sobre agricultura. Por que, então, alguém iria procurar sob esse termo? Existem várias possibilidades:

1. descobrir quais os filmes que têm locação numa fazenda,
2. contar quantos desses filmes já foram realizados, a fim de estudar tendências da produção cinematográfica durante certos períodos, ou
3. identificar o título de um filme quando o máximo que a gente se lembra é que ele se passava numa fazenda ou numa comunidade de agricultores.

A segunda dessas possibilidades sugere alguma aplicação para fins acadêmicos. A primeira sugere uma forma de uso em 'produção' (por exemplo, produtores de cinema ou televisão que procuram descobrir como um acontecimento, lugar, pessoa, atividade ou profissão foi representado por outros). O terceiro tipo de questão representa um uso mais popular. Trata-se, no entanto, do tipo de consulta que é quase certo aparecer no departamento de referência de uma biblioteca pública. Na medida em que surgem questões dessę tipo, parece inteiramente justificável alguma forma de indexação temática de filmes de ficção, mesmo que estes não sejam realmente considerados como se 'tratassem do' conteúdo temático implícito nos termos de indexação. Exatamente o mesmo argumento pode ser levantado em relação a alguma forma de 'abordagem temática' da literatura de ficção. O romance *20 000 léguas submarinas* pouco contribui, se é que contribui, para nosso conhecimento sobre submarinos. É improvável que alguém considere sensatamente que esse romance 'trata de' submarinos. No entanto, a indexação temática da literatura de ficção tem sua utilidade. Alguém pode legitimamente querer saber 'quais os romances passados em submarinos?", 'quantos romances se passaram em submarinos?', 'qual foi o primeiro romance que aconteceu num submarino?' ou 'qual foi aquela obra antiga famosa que previu o desenvolvimento de submarinos?'

Isso não deve ser encarado como se estivéssemos a sugerir que filmes e romances não têm substância, não têm atinência. O filme *Patton — rebelde ou herói?* trata evidentemente do general Patton. Continuaria tratando de Patton mesmo que contivesse pouca exatidão histórica. A maioria dos espectadores concordaria em que o filme mostra a ambição de Patton. Se isso faz com que o filme trate de ambição ou justifique sua indexação sob o termo AMBIÇÃO é uma questão inteiramente diferente. O filme também mostra a guerra travada com tanques. Isso significa que ele trata de guerras com tanques? Ele trata de generais, de estratégia militar? Pode-se considerá-lo como se tratasse da Inglaterra ou da França só porque partes do filme se passam nesses dois países?

De um ponto de vista prático, evidentemente, o que estamos examinando é, de fato, uma relação entre uma obra e os termos de indexação atribuídos para representar essa obra. Quando atribuímos um termo de indexação a um livro ou artigo de periódico estamos admitindo, em quase todos os casos, que a obra veicula alguma informação sobre o tópico representado por esse termo. No caso de uma obra de ficção, por outro lado, pode-se atribuir-lhe um termo de indexação por outros motivos, principalmente para representar:

1. Seu tema central ou temas.
2. O que ela pode exemplificar, talvez casualmente.
3. O ambiente em que ela se situa.

Na realidade, os dois primeiros motivos acima não são, é claro, significativamente diferentes. Pelo menos, a única diferença diz respeito à extensão com que o tema é tratado.

O ambiente do filme pode ter dimensões espaciais, temporais e de 'personagens'. A dimensão espacial pode ser bastante precisa — Pigalle, Paris ou França — ou imprecisa — uma floresta, um rio, uma comunidade rural. A dimensão temporal, igualmente, pode ser precisa — por exemplo, a Revolução Francesa — ou maldefinida (por exemplo, o século XIX ou 'antes de Cristo'). A dimensão de 'personagens' refere-se ao ambiente criado pelos tipos de personagens representados. O fato de o personagem principal de um filme ou romance ser uma enfermeira não faz com que tratem necessariamente de enfermagem ou mesmo de enfermeiras. No filme *Doutor Jivago*, Lara aparece em diferentes momentos como estudante universitária, enfermeira e bibliotecária. O filme, de fato, não trata de nenhum desses papéis, pois dificilmente seria possível vê-lo como um repositório de informações sobre eles. Por outro lado, em certo sentido, *Nunca te amei* (*The Browning version*) pode ser visto como uma obra que trata de mestres e ensino, pois as relações entre professor e aluna são fundamentais no enredo. O ensino não é simplesmente um 'ornamento' ou uma imposição do ambiente.

De um ponto de vista pragmático nada disso é realmente importante. A questão fundamental não é se uma obra trata de enfermagem, utiliza a enfermagem como exemplo ou ambiente, mas se o termo de indexação ENFERMAGEM lhe deve ser aplicado.

Uma das grandes diferenças entre a indexação de obras de ficção e a indexação de outros tipos de trabalhos é que as primeiras provavelmente são mais subjetivas e interpretativas do que as outras. Estudos sobre coerência da indexação demonstraram que não é provável que diferentes indexadores concordem inteiramente sobre quais termos devam ser atribuídos a determinado item, mesmo quando o conteúdo temático de que ele trata seja razoavelmente concreto. No caso de obras de ficção, provavelmente é muito menor a possibilidade de que venha a existir concordância. Isso seria especialmente verdadeiro no caso em que a obra de ficção trata fundamentalmente de alguma emoção ou qualidade — ciúme, ambição ou cobiça, por exemplo.

As obras de ficção possuem outra característica importante que complica a indexação temática: seu campo de ação é essencialmente aberto. Isto é, não há limites de fato para aquilo que podem representar. Neste sentido, a indexação dessas obras tem algo em comum com a catalogação de assuntos numa grande biblioteca geral ou a indexação de um jornal de conteúdo genérico. Pelo menos tem mais em comum com isso do que com um ambiente de assuntos mais restrito, como a indexação de uma coleção de itens em agricultura ou educação. O vocabulário usado na indexação deve também ser aberto, pois constantemente estão sendo feitos filmes e romances que tratam de personalidades, acontecimentos e lugares que não foram anteriormente abordados por esses meios.

Duas importantes considerações relativas à indexação de assuntos são:

1. quem deve fazer a indexação, e
2. a quais diretrizes os indexadores devem obedecer.

Na indexação de obras especializadas em uma área de assunto delimitada, é evidente que se torna necessário algum nível de conhecimento especializado. O grau desse conhecimento especializado que será necessário dependerá, em grande medida, do grau de hermetismo do conteúdo temático e de sua terminologia. Percebe-se intuitivamente que a indexação em matemática ou mecânica aplicada talvez exija maior domínio do assunto do que a indexação, por exemplo, na área de transportes, cuja terminologia apresenta maior probabilidade de ser conhecida do público em geral. Um bom indexador não precisa necessariamente ser um especialista num assunto; inversamente, um especialista num assunto não faz necessariamente um bom indexador.

Uma vez que o conteúdo de obras de ficção não é limitado por sua temática, nessa situação o domínio de um assunto, no sentido convencional, é irrelevante. Ademais, pode-se considerar que aquilo que essas obras apresentam cai na categoria de 'conhecimentos gerais' e nada tem a ver diretamente com as técnicas envolvidas na produção de obras de ficção. Não há motivo algum para supor, por exemplo, que a indexação de filmes deva ser feita por estudiosos de cinema (embora essas pessoas possam oferecer contribuições valiosas a respeito dos tipos de termos que seriam úteis, pelo menos para elas) ou mesmo que ela exija algum conhecimento específico das técnicas da cinematografia.

Duas características de um índice que terão importante impacto em seu desempenho são:

1. a exaustividade da indexação, e
2. a especificidade dos termos utilizados.

Conforme vimos antes neste livro, a *exaustividade* refere-se à extensão com que o conteúdo de uma obra é coberto pelos termos utilizados na indexação. A exaustividade diz respeito ao âmbito de cobertura. Neste contexto, o oposto de 'exaustivo' é 'seletivo'. Em geral, a exaustividade equivale ao número de termos de indexação utilizados. Se o filme *Geronimo* aparecesse, num índice, somente sob os termos GERONIMO e ÍNDIOS APACHES, essa indexação seria bastante seletiva. No índice de assuntos da primeira edição do *American Film Institute Catalog*, este filme, no entanto, foi indexado sob 17 termos diferentes; trata-se de uma indexação bastante exaustiva.

Há prós e contras na alta exaustividade. Em teoria, a indexação exaustiva facilita achar as coisas: a possibilidade de localizar um item crescerá quase com certeza à medida que crescer o número de pontos de acesso (isto é, entradas). Isso, porém, só é verdade até certo ponto. Se a indexação for excessivamente exaustiva causará uma diluição da eficácia do índice — o menos importante ocultará o mais importante e dificultará sua localização. Num exemplo extremo, será muito difícil identificar filmes ou romances que tratem, com algum interesse, de cães, se o termo de indexação CÃES for atribuído a toda obra em que apareça um cão, mesmo de passagem. A questão, evidentemente, é que a indexação de assuntos comumente implica uma decisão binária simples (um termo é ou não é aplicado) e não uma decisão ponderada (um termo se aplica com certo peso). Por isso, em certas aplicações do índice, o joio pode ocultar o trigo.

Enquanto a exaustividade diz respeito ao âmbito de cobertura, a *especificidade* refere-se à profundidade de tratamento do conteúdo. Estabelece-se a exaustividade como uma decisão da política de indexação, enquanto a especificidade é uma propriedade do vocabulário adotado na indexação. Em geral, constitui uma boa prática de indexação empregar o termo mais específico disponível para descrever algum aspecto presente no documento. Este princípio, entretanto, precisa ser temperado com o senso comum. Ao projetar um índice, deve-se tentar chegar a um nível de especificidade que seja apropriado às necessidades dos usuários desse índice. Lassie é presumivelmente uma cadela da raça cole. Seria tecnicamente correto indexar os filmes de Lassie sob COLES. No entanto, percebe-se intuitivamente ser improvável que os usuários de um catálogo de filmes precisem, ou procurem, qualquer termo mais específico do que CÃES. Por outro lado, seria preciso indexar de modo muito mais específico do que CÃES numa enciclopédia sobre animais de estimação. Evidentemente, quanto mais específicos forem os termos utilizados, menos entradas por termo haverá em média. Isso facilita a localização de algo altamente específico, mas torna mais difícil a realização de buscas mais genéricas.

A ficção em particular

Embora a maioria dos exemplos usados até agora estivesse relacionada a filmes, os mesmos argumentos e princípios são aplicáveis a romances e outras obras de ficção em formato impresso. Apesar de alguns autores, notadamente Pejtersen (ver Pejtersen, 1979, 1984); Pejtersen e Austin, 1983, 1984) terem realizado experimentos com a indexação de ficção, ao longo de muitos anos, o interesse pelo tema aumentou notavelmente na última década, a ponto de ter levado a American Library Association a publicar 'diretrizes' sobre a questão (*Guidelines on subject access*, 2000).

Pejtersen (1992), entre outros, chamou atenção para a anomalia relativa ao fato de que os bibliotecários geralmente pouco fizeram para melhorar o acesso às obras de ficção, muito embora elas representem a metade do acervo das bibliotecas públicas e mais da metade das que são retiradas por empréstimo.

Sapp (1986) e Baker e Shepherd (1987) estudam a classificação de obras de ficção nas estantes das bibliotecas e as limitações dos esquemas de classificação bibliográfica ou das listas de cabeçalhos de assuntos existentes, que pouco contemplam o acesso temático às obras de criação. Baker (1988) descreve os resultados de experiências com a classificação de obras de ficção em bibliotecas públicas. Sapp (1986) também examina os métodos adotados em certas fontes impressas, como o *Short Story Index*, o *Cumulated Fiction Index* e o *Fiction Catalog*. Embora essas publicações realmente indexem os enredos sob mais de um cabeçalho, padecem das desvantagens dos índices impressos em geral — não permitem ao usuário combinar cabeçalhos numa busca. Assim, seria possível identificar histórias policiais e histórias que se passam na China, mas seria muito mais difícil identificar histórias policiais que se passam num ambiente chinês.

Olderr (1991) salientou por que a indexação de obras de ficção é importante para as bibliotecas:

> Nunca é fácil responder a perguntas do tipo 'você tem algum romance policial que se passe em Iowa?' ou 'existe algum romance atual sobre a morte? ou 'você poderia me sugerir um romance sobre o esforço de guerra em território inglês durante a Segunda Guerra Mundial?' (p. xiii).

Guard (1991) também analisa as formas de abordar a ficção de que precisam os usuários típicos de uma biblioteca, e Hayes (1992b) apresenta os resultados de algumas experiências sobre 'acesso melhorado ao catálogo' de obras de ficção em bibliotecas, detendo-se principalmente no tempo destinado à catalogação e nos tipos de cabeçalhos necessários. Ranta (1991) apresenta uma perspectiva diferente, argumentando que o acesso temático a obras de ficção é necessário para facilitar várias modalidades de estudos literários.

Um método avançado para indexação de literatura de ficção foi descrito por Pejtersen (por exemplo, 1979, 1984) e Pejtersen e Austin (1983, 1984). Baseando-se numa análise sobre como os usuários de bibliotecas públicas caracterizam o conteúdo dos livros, Pejtersen identificou quatro 'dimensões' principais da

obra de ficção: conteúdo temático, referencial (época, lugar, meio social, profissão), intenção ou atitude do autor, e acessibilidade. A partir disso, ela criou um esquema de indexação que envolvia as seguintes dimensões e categorias:

1. Conteúdo temático
   a. ação e curso dos acontecimentos
   b. desenvolvimento e descrição psicológica
   c. relações sociais
2. Referencial
   a. época: passado, presente, futuro
   b. lugar: geográfico, meio social, profissão
3. Intenção do autor
   a. experiência emocional
   b. cognição e informação
4. Acessibilidade
   a. legibilidade
   b. características físicas
   c. forma literária

O esquema foi adotado, na Dinamarca, na indexação de várias bases de dados em linha, e mais recentemente no catálogo interativo em linha conhecido como Book House. Permite fazer buscas a partir de dados bibliográficos, palavras-chave controladas, termos de classificação, e palavras/expressões constantes de uma anotação em linguagem natural. A figura 93 (extraída de Pejtersen, 1992) mostra uma entrada completa do Book House. A figura 94 é um exemplo anterior, com a indexação completa de um romance por meio de palavras-chave.

Autor:	Haller, Bent
Título:	Kaskelotternes sang, 1983, 137 páginas
Capa:	Azul, mar, baleias, *icebergs*
Nomes:	Tangeje, Peter
Conteúdo temático:	A vida no mar de um filhote de cachalote. Sua luta pela sobrevivência apesar da poluição, da fome e da matança das baleias pelo homem. A união dos cachalotes na luta contra os perigos do mar.
Ambiente:	Ambiente marinho.
Época:	Década de 1980.
Cognição:	Crítica à poluição dos mares pelo homem e à matança das baleias, levando-as quase à extinção.
Experiência emocional:	Emocionante, triste.
Forma literária:	Romance, história de animais.
Legibilidade:	11 anos de idade, leitura em voz alta para crianças a partir de 7 anos (final feliz).
Tipografia:	Letras graúdas.

FIGURA 93
Exemplo de entrada da base de dados de ficção Book House
Reproduzido de Pejtersen (1992) com permissão de Emerald

BRANNER, H.C. *Barnet leger ved stranden*

Descrição psicológica: Depois de um casamento fracassado, um homem se isola em um chalé de veraneio, vivendo uma profunda crise. Encontra duas pessoas, que sobre ele exercem influência. *Época*: década de 1930. *Lugar*: Dinamarca, um chalé de veraneio à beira-mar. *Ambiente social*: Classes médias. *Cognição/informação*: a relação entre as experiências da infância e os medos e fracassos matrimoniais da vida adulta. Perspectiva psicoanalítica. *Legibilidade:* Difícil. *Tipos usados na composição:* Graúdos. *Forma:* Diário. *Dados bibliogr.:* Copenhague: Povl Branner, 1937.–379 p.

Pontos de acesso: 1930–1939
Diários
Depressão
Medo
Culpa
Descrições psicológicas
Problemas de identidade
Problemas psicológicos
Repressão

FIGURA 94
Exemplo de um romance indexado com o emprego do método de Pejtersen
Reproduzido de Pejtersen e Austin (1983) com permissão de Emerald

Uma das principais vantagens de um método tão estruturado quanto esse para a indexação da literatura de ficção é que permite que sejam realizadas buscas adotando-se uma espécie de modo de 'comparação de padrões', que serve para muitos leitores que desejam livros 'similares' a um que tenham lido recentemente. Os critérios pelos quais as obras de ficção são procuradas pelos usuários de bibliotecas são mais pessoais e idiossincráticos do que os critérios e as características comumente associados às buscas por assuntos em bases de dados bibliográficos que abranjam, por exemplo, artigos de periódicos. Embora isso apresente importantes desafios para quem projeta sistemas de recuperação, também sugere enfoques inovadores do problema da recuperação da informação. Imagine-se uma base de dados de biblioteca pública que armazenasse informações sobre as obras de ficção retiradas por empréstimo por cada cliente. Seriam, então, desenvolvidos programas que identificariam grupos (talvez pares) de clientes que tivessem muitos livros em comum. Essa informação, em seguida, seria utilizada para gerar listas de sugestões de leitura para os usuários da biblioteca. Por exemplo, se o Usuário *A* tomou emprestados os itens *a, b, c, d, e,* e o Usuário *B* tomou emprestados *a, d, e* e *f*, talvez *A* viesse a se interessar pela existência de *f* e *B* pela existência de *b* e *c*. O sistema de Pejtersen permite, de fato, realizar buscas por 'livro-modelo', isto é, localizar um romance 'similar' a outro que foi considerado divertido. 'Similar' poderia ser em termos de cenário, tema, ponto de vista do autor, experiência emocional, e assim por diante.

Beghtol (1994) é um tanto crítico do esquema de indexação de Pejtersen, revindicando melhores resultados para uma classificação alternativa e muito

anterior (Walker, 1958), embora baseada na análise detalhada de um único romance, e propondo um esquema minucioso de sua própria autoria, que empregava um método de classificação facetada.

A publicação da American Library Association sobre indexação de obras de criação (*Guidelines on subject access*, 2000) é menos uma série de diretrizes do que um vocabulário, em formato de tesauro, porém baseado nos cabeçalhos de assuntos da Library of Congress, que pode ser empregado para indexar ficção, peças de teatro e outros gêneros. O vocabulário abrange apenas tipos de obras (por exemplo, poesia histórica, filmes de horror, romances históricos); os usuários são encaminhados a outras fontes, a fim de verificar a forma correta dos nomes dos personagens, dos nomes de lugares e outros pontos de acesso.

As diretrizes que realmente aparecem na publicação da ALA são bastante imprecisas. Além de termos para formas, as diretrizes contemplam a atribuição de termos para personagens, ambientes e 'tópicos'. O ambiente refere-se tanto a lugares quanto a períodos, e devem ser adotados subcabeçalhos de forma (por exemplo, Paris (França) – Poesia). As diretrizes especificam que os nomes de personagens fictícios e lendários (ao contrário de pessoas reais) somente devem ser usados "quando surgirem com destaque em três ou mais obras". Embora um indexador relativamente culto provavelmente saiba que Sherlock Holmes e Narnia aparecem em muitas obras, como poderia alguém saber que um detetive ou um lugar menos famosos se encontram em pelo menos três obras, a menos que esse alguém tivesse à mão várias dessas obras imediatamente. E, além do mais, o que há de tão especial no número 'três'?

As diretrizes da ALA sobre 'acesso tópico' são ainda mais vagas:

> Atribua tantos cabeçalhos tópicos quantos forem justificáveis pelos assuntos da obra. As sobrecapas dos livros e as recensões são uma boa fonte de informação para identificar de que trata uma obra. Caso não existam, uma técnica muitas vezes eficiente é 'passar os olhos' no texto para identificar seu conteúdo tópico.
>
> Os temas de obras de ficção, identificados na crítica literária, podem ser expressos com cabeçalhos dos *LCSH* representativos de qualidades ou conceitos. Uma vez, porém, que os *LCSH* foram projetados para indexação de obras que não são ficcionais, são comparativamente poucos os cabeçalhos que se prestam a tal fim (p. 47).

De fato, o folheto da ALA não serve a nenhum propósito útil, pois as diretrizes são muito vagas e há um tesauro mais completo e melhor (Olderr, 1991).

Em novembro de 1991, o OCLC e a Library of Congress deram início a uma experiência de catalogação cooperativa de assuntos em textos de ficção, dramaturgia e outras obras de criação. Várias bibliotecas públicas e universitárias participaram do OCLC/LC Fiction Project contribuindo para a complementação de registros MARC de um conjunto de itens selecionados. Foram a eles acrescentados termos relativos tanto a gênero quanto a assunto (cabeçalhos de assuntos LC). Mais de 15 000 registros LCMARC foram complementados pelo OCLC e as bibliotecas participantes. Além disso, foram também complementados registros

bibliográficos feitos por algumas das bibliotecas participantes, e muitas propostas de cabeçalhos de assuntos foram submetidas à Library of Congress, que aprovou mais de mil dessas propostas, em sua maioria cabeçalhos para personagens de ficção (Westberg, 1997). O projeto foi concluído em 1999.

Em 1997, a *British National Bibliography* passou a incluir entradas de obras de ficção com cabeçalhos de assuntos tópicos, bem como cabeçalhos de gênero e forma baseados nas diretrizes da ALA (MacEwan, 1997).

É provável que as obras de ficção apresentem dificuldades maiores para o indexador do que outros tipos de publicações. A coerência provavelmente será até menor, a menos que seja adotado um vocabulário controlado de termos genéricos, bem pequeno, principalmente se o indexador tiver de expressar o 'ponto de vista' do autor. A indexação da literatura de ficção (por exemplo) parece inerentemente mais subjetiva do que a indexação de periódicos ou livros especializados que tratam de ficção. Outro problema é que não é absolutamente fácil, para os objetivos da indexação, fazer a leitura por alto de obras de ficção, e o indexador não conta com o auxílio dos títulos e entretítulos temáticos, que quase certamente encontra em muitos outros tipos de publicações (Jonak, 1978).

Olderr (1991) identifica os problemas com bastante clareza:

> A catalogação de obras de ficção exige imaginação. Uma obra de não-ficção, mesmo que não traga dados de Catalogação na Publicação (CIP) no verso da folha de rosto, possui um sumário, um índice, títulos temáticos dos capítulos e outras características que ajudarão o catalogador. Até o título normalmente reflete com precisão o conteúdo. Se o livro for sobre a inveja, assim haverá de declarar; se for sobre ciúme, também o dirá. Uma obra de ficção, por outro lado, pode tratar da inveja ou do ciúme e jamais empregar no texto uma dessas palavras. E depois que o catalogador houver identificado o tema, ainda haverá o problema de lembrar qual é a diferença entre inveja e ciúme. Isso, para começar, não é algo que seja do pleno conhecimento de todos... (p. xiv).

DeZelar-Tiedman (1996) estudou a factibilidade de empregar informações fornecidas pela editora (por exemplo, as constantes da sobrecapa ou da capa) como fonte de termos representativos de personagens, ambiente, gênero e tópico. Em geral, ela considerou que isso era satisfatório para a maioria dos itens, porém a amostra em que se baseou era muito pequena.

Down (1995) examina alguns dos problemas com que ela se defrontou na atribuição de cabeçalhos de assuntos a obras de ficção. Sua experiência sugere ser improvável que o exame superficial de um romance ou a confiança nas informações fornecidas pela editora possam esclarecer quais sejam realmente os temas que a obra ilustra.

Beghtol (1994) oferece o levantamento mais completo dos problemas da indexação de obras de ficção, inclusive a questão da 'atinência', além de apresentar seu próprio método.

Nielsen (1997), recorrendo ao campo da crítica literária e dos estudos literários, argumenta que a indexação e redação de resumos de ficção constitui uma forma de interpretação literária. Afirma que as abordagens da indexação de ficção, inclusive a de Pejtersen, concentram-se no *quê* trata o livro e pouca atenção dedicam a *como* a história é contada. Ele menciona alguns elementos, como o estilo, a narrativa, o modo discursivo e a composição, como alguns dos elementos do aspecto relativo ao *como* da ficção.

Nielsen oferece maiores informações sobre quais os tipos de coisas a serem considerados na indexação do aspecto relativo ao *como* de um romance:

> *-Gênero, subgênero, tipo literário.* (Qual o tipo de literatura?)
>
> *-Estrutura narrativa, enredo.* (Por exemplo, trata-se de uma estrutura simples ou complexa? Uma estrutura linear, cronológica, ou uma alternância entre tempos diferentes? Ou a estrutura é formada por variações de fragmentos, colagem, não cronológica mas tematicamente organizada? A narrativa é estruturada como um quebra-cabeça?)
>
> *-A maneira de contar do(s) narrador(es).* (Por exemplo, como a narrativa é apresentada? Quantos narradores? O narrador fala na primeira ou na terceira pessoa? Narrador distanciado ou comprometido? 'Mostra' ou 'conta'?)
>
> *-Pontos de vista.* (Por exemplo, a história é contada a partir de um ponto de vista específico? Ou há uma alternância entre diferentes pontos de vista?
>
> *-Estilo, maneira de contar, estrutura do discurso.* (Por exemplo, estilo específico: impressionista, surrealista, etc. Mais genérico: maneira didática, cômica, irônica de contar; discurso que usa a linguagem corrente, ou que usa trocadilhos, estrutura ilógica do discurso, ou alternação entre os discursos mais diferentes; intertextualidade.)
>
> *-Função do ambiente.* (A função é documentária? É a convencional para esse tipo específico de romance? Ou o ambiente é empregado de forma simbólica ou alegórica?)
>
> *-Padrões de metáforas, motivos determinantes, simbolismo.* (O simbolismo é discreto ou dominante? Quais os tipos de símbolos utilizados? Qual o tipo de motivo determinante que pode ser encontrado? Quais os símbolos, motivos, alegorias ali encontrados? Por exemplo, o motivo do duplo de alguém, o motivo de Don Juan, o mito do Paraíso) (p. 174-175).

Embora a indexação desses aspectos revista-se de utilidade para os estudiosos da literatura, é improvável que venha a ter muito interesse para os leitores típicos de obras de ficção. Ademais, esse tipo de indexação exigiria uma análise textual minuciosa que somente um especialista em literatura poderia proporcionar. Isso seria inutilmente dispendioso em qualquer aplicação que tivesse uma dimensão significativa.

Trabalhos sobre indexação e resumos de obras de ficção são também objeto de uma série de artigos de Saarti (1999, 2000a,b, 2002). Um estudo sobre coerência de indexação foi realizado em cinco bibliotecas públicas finlandesas. Cinco romances iguais foram indexados por três bibliotecários e três usuários de cada biblioteca. Os termos foram extraídos de um tesauro finlandês para indexação de ficção e os indexadores foram solicitados a redigir resumos dos romances antes de indexá-los. Obviamente, a coerência foi baixa e houve variações muito grandes de um indexador para outro quanto ao número de termos atribuídos. Os indexadores bibliotecários atribuíram menos termos do que os usuários e foram mais coerentes entre si. No entanto, o valor de sua coerência foi de apenas 19,9% em comparação com 12,4% dos usuários. Os romances mais 'complexos' (por exemplo, os de Dostoievski) foram indexados com mais termos do que os menos complexos (por exemplo, de Simenon). Os resumos variaram de

tamanho de 23 a 186 palavras (média de 68). Cerca de 75% dos 3 206 diferentes 'elementos' dos resumos lidavam com conteúdo (como temas, ambientes e personagens), 11,9%, com a estrutura do romance, 5,5%, com a experiência subjetiva da leitura, e 5,2%, com a crítica ou avaliação do romance. Os usuários foram mais avaliadores/críticos do que os bibliotecários (Saarti, 2000a,b). Saarti (1999) trata de tesauros para a indexação de ficção e, em particular, do tesauro finlandês.

Este exame da questão partiu da hipótese de que as obras ficcionais são indexadas em alguma forma de base de dados. Bradley (1989) examina uma situação afim a essa: a necessidade de índices nas próprias obras de ficção. Embora seja defensável a inclusão de índices no final de certas obras, como, por exemplo, clássicos renomados, certos romances históricos e outros trabalhos ficcionais que possam ser objeto de pesquisas científicas, o estudo de Bradley mostrou que foi pouco o interesse demonstrado por romancistas, críticos, leitores ou editoras.

Bell (1991b) identifica os problemas especiais implícitos na elaboração de índices de romances. Ela salienta que dar a entender a 'sutileza e complexidade' da intenção de um romancista é muito mais difícil do que expressar de que trata uma obra de não-ficção.

Redação de resumos

As obras de ficção, tanto quanto outros tipos de publicações, precisam ser resumidas (quando não seja, para facilitar sua indexação), porém as características dos respectivos *resumos* ou *sinopses* são bastante diferentes das características dos resumos de publicações científicas examinados anteriormente neste livro. Um bom resumo deve conter os aspectos fundamentais do enredo ou ação, indicando o ambiente (geográfico, cronológico) e as emoções descritas, quando isto for apropriado. A sinopse pode ser estruturada como no exemplo da figura 93 ou adotar a forma de uma narrativa simples, como no exemplo da figura 95. Embora as características da sinopse sejam bastante diferentes das características do resumo, sua finalidade principal é semelhante — indicar para o leitor se ele precisa ou não ler ou ver o item descrito. Além disso, aplicam-se igualmente à sumarização de obras de ficção os mesmos princípios básicos que orientam a redação de resumos: exatidão, brevidade, clareza.

> Um coelho invade uma horta para comer as hortaliças. O hortelão descobre-o e passa a persegui-lo. O coelho foge.
>
> Pedro, o Coelho, imprudentemente, invade o canteiro de uma horta para comer as hortaliças. O dono, Sr. McGregor, descobre-o e procura livrar sua horta desse animal pernicioso. Depois de uma perseguição angustiante, Pedro consegue fugir e voltar para casa.

FIGURA 95
Duas sinopses possíveis de *As aventuras de Pedro, o Coelho,* de Beatrix Potter
Apud Krieger (1981), com modificações, e com permissão do autor

São muito poucas as diretrizes existentes sobre preparação de sinopses de literatura de ficção. A editora de *Masterplots* (Magill, 1976) oferece alguma orientação, mas de uma forma muito geral:

> Projetado fundamentalmente para consulta, o formato de MASTERPLOTS é estruturado e padronizado, a fim de oferecer o máximo de informação da forma mais rápida. Cada uma das sínteses é precedida de dados de referência cuidadosamente verificados e enunciados sucintamente, os quais informam num relance o tipo de obra, autoria, tipo de enredo, época do enredo, lugar e data da primeira edição. Em seguida encontra-se uma lista dos personagens principais e as relações entre eles, o que muitas vezes é uma característica bastante útil. Depois vem a *Crítica*, uma análise breve e incisiva do livro original. Finalmente segue-se o resumo do enredo, apresentado como uma história completa e isenta de citações da obra original (p. v).

Em *Masterplots II* (Magill, 1986) foi adotado um formato um pouco diferente:

> [...] junto com uma síntese do enredo, com freqüência se examinam os recursos narrativos e se estuda a construção dos personagens de forma mais profunda do que antes — um aspecto que é útil para os estudantes mais jovens. Além disso, identificam-se e se analisam os principais temas do romance em questão, e o êxito em geral dos esforços do autor é comumente analisado num resumo interpretativo (p. vii-viii).

A figura 96 apresenta um exemplo de *Masterplots II.*

A edição revista de *Masterplots II* (Kellman, 2000) adota uma abordagem mais estruturada da sinopse do enredo (conceitualmente similar a um resumo estruturado) com quatro componentes: Enredo, Personagens, Temas e Significados, e Contexto Crítico. Seu emprego é descrito da seguinte forma:

> Esta análise começa com um resumo dos principais elementos do enredo da obra e continua com seções separadas que a examinam em profundidade. A seção 'Os Personagens' examina as motivações e o desenvolvimento das pessoas retratadas; 'Temas e Significados' examina as preocupações maiores da obra; e 'Contexto Crítico' avalia o lugar da obra na tradição literária norte-americana e sintetiza qual foi sua recepção. Cada verbete termina com uma bibliografia comentada que orienta o leitor para outras fontes recentes de estudo (p. v).

Pejtersen (1994) admite três estruturas lingüísticas básicas para identificar e expressar o conteúdo de obras ficcionais (ver figura 97). Esse esquema pode ser empregado para orientar a redação de anotações, como no exemplo apresentado, e essas anotações são uma fonte óbvia de termos de indexação úteis. Ela adverte, no entanto, que "uma descrição completa do conteúdo temático pode exigir a combinação de várias estruturas".

A BLOODSMOOR ROMANCE

Author: Joyce Carol Oates (1938-)
Type of plot: Historical romance fantasy
Time of plot: 1879-1900
Locale: Bloodsmoor, a valley in Eastern Pennsylvania
First published: 1982

Principal characters:
- JOHN QUINCEY ZINN, a gentleman-inventor and the father of a large family
- PRUDENCE KIDDEMASTER ZINN, his wife, mother of the Zinn daughters
- CONSTANCE PHILIPPA, their oldest daughter who later becomes a son
- MALVINIA, another daughter, later a famous actress
- OCTAVIA, another daughter, later a wife and mother
- SAMANTHA, another daughter who serves as her father's laboratory assistant
- DEIRDRE, an adopted daughter and spiritualist

The Novel

Joyce Carol Oates's book *A Bloodsmoor Romance* is not a kind of fiction that is easily named, although it is not hard to recognize. The work combines both realism and fantasy in a display of authorial skill: Oates uses several techniques to achieve this effect. First, she sets her romance in a past that closely resembles the historical past; in that setting one finds both fictional characters and characters who bear the names of figures from history. In addition, the characters of the work are interested in many of the things that interested the real nineteenth century: spiritualism, the theater, the westward movement, experimental science, abnormal psychology, female sexuality, and the nature of marriage.

It is Oates's second technique that sets the work apart from historical romances per se: She freely manipulates the order of historical events and even adds events that could not possibly occur. John Quincey Zinn demonstrates both of these intrusions of fantasy: He invents the ballpoint pen and solar heating but dismisses them as useless. He invents an operating time machine, but he destroys it after he uses it to misplace one of his pupils. Similarly, Zinn's daughter Constance combines fantasy with history. Reared for marriage, Constance spends her early life accumulating household linens, but when the wedding night comes, she panics, and placing in her groom's bed the dress form used to fit her trousseau, she runs away. Disguising herself as a man, she heads west and tries her hand at being a

FIGURA 96
Exemplo de uma entrada de *Masterplots II* (1986)

cowboy, an outlaw, a deputy sheriff, and a gambler. During her masquerade, she turns physically into a man as well, and when she returns to the family home at Bloodsmoor, she poses as Philippe Fox, Constance's agent. Eventually, "he" apparently elopes with a childhood girlfriend.

The plot of the book unfolds by following the lives of the daughters as they grow up. In their adventures, the reader meets several characters drawn from history. For example, Deirdre, the Zinns' adopted daughter, is kidnaped by a mysterious stranger in a black balloon who deposits her on the lawn of a character named Madame Elena Blavatsky. This Madame Blavatsky shares the quirks of the historical Madame Blavatsky, cofounder of the American Theosophical Society. Recognizing Deirdre's talents, Oates's Blavatsky teaches Deirdre to become a medium, contacting spirits beyond the grave, and takes her on a world tour. The reader meets other fictional characters with real counterparts as well: Mark Twain, for one.

As may be inferred from the events recounted above, *A Bloodsmoor Romance* is an often hilariously comic work, yet one that at the same time attempts to capture some of the boundless enthusiasm of the late nineteenth century, an enthusiasm that was often as undiscriminating as it was energetic.

FIGURA 96 (continuação)
Exemplo de uma entrada de *Masterplots II* (1986)

Estrutura 1:

Personagem (ns) principal(is) como substantivo no genitivo — acontecimento central como substantivo — elementos de sujeitos remanescentes como orações prepositivas.

Exemplo: A carreira militar de um tenente inglês. Suas perigosas expedições contra contrabandistas e seu trabalho como capitão de um navio corsário.

Foco: Personagem principal.

Estrutura 2:

Acontecimento(s) central(is) como substantivo — elementos sujeitos como orações prepositivas.

Exemplo: Tráfico de drogas entre um agente da CIA em missão secreta e a embaixada chinesa.

Foco: Acontecimento central.

Estrutura 3:

Personagens principais como substantivo — estrutura 1 — elementos de sujeitos remanescentes como orações prepositivas.

Exemplo: A prisão de Hall. O cotidiano dos prisioneiros, seu vício em drogas, suas provocações e revolta contra os funcionários da penitenciária.

Foco: Relações entre personagens e acontecimentos principais.

FIGURA 97
Estruturas lingüísticas para orientar a anotação e indexação de ficção

CAPÍTULO 13

Bases de dados de imagens e sons

Um livro organizado por Feinberg (1983) examina várias questões especiais da indexação, mas se limita quase que exclusivamente à indexação de textos impressos em papel. Toda área do conhecimento, bem como distintos formatos impressos, como jornais e leis, suscita problemas de indexação algo diferentes. As diferenças de indexação presentes nessas variantes são, porém, de somenos. Mais relevantes são as questões que surgem ao sairmos do texto impresso para outros formatos. Este capítulo examina a indexação de imagens e sons gravados. São áreas difíceis, pois abarcam campos, como tecnologia da fala, visão computacional, e compreensão de documentos, que ultrapassam em muito o escopo da maioria das aplicações da indexação.

Indexação de imagens

A capacidade de armazenar, em formato digital, em bases de dados, qualquer tipo de imagem, e especialmente de poder acessar milhões delas na Rede, causou impressionante ressurgimento do interesse por imagens em geral e, em particular, por modos de indexá-las. Disse Jörgensen (2001) sobre essa revolução:

> Encontramo-nos, ao que parece, no ponto crítico de importante movimento histórico de retorno ao que se poderia chamar o primado da imagem. Ao longo dos últimos séculos, as palavras foram a forma privilegiada de comunicação e o meio preferido de educação. Uma mudança, porém, se verificou nas últimas décadas, e as imagens vêm reafirmando sua primazia como mensageiros instantâneos e poderosos (p. 906).

Tudo que foi dito sobre indexação neste livro, até aqui, limitou-se a textos escritos. É claro que descrever imagens com palavras ainda é importante. Imagens digitais, porém, também podem ser indexadas (automaticamente) e recuperadas por atributos intrínsecos, como cor, forma e textura. Os termos que distinguem os dois métodos não são de todo coerentes, mas a descrição de imagens, com palavras, feita por seres humanos, denomina-se em geral indexação *baseada em conceitos*, e a indexação de imagens por seus atributos intrínsecos é *baseada em conteúdos* (Rasmussen, 1997). Características como cor, forma e textura são amiúde denominadas *características de nível baixo*. As *características de nível alto* são descrições da imagem baseadas em palavras.*

* Alguns autores, como Mostafa (1994), distinguem entre indexação verbal (isto é, representação textual de uma imagem) e indexação baseada em imagens (a extração de características, e, portanto, pontos de acesso, da própria imagem), e isso parece ser uma diferenciação clara, exceto, naturalmente, que um único sistema de recuperação pode incluir ambos os tipos.

Besser (1997) chamou a atenção para o problema da indexação relativa a imagens da seguinte forma:

> Como as coleções de imagens possuem muito poucas informações textuais que originalmente as acompanhem, nossos sistemas tradicionais de recuperação não se aplicam facilmente a elas [...] Os museus, que, coletivamente, abrigam um dos maiores conjuntos de imagens que efetivamente vêm acompanhadas de texto, muitas vezes atribuem termos a uma imagem que não são absolutamente úteis para o leigo (p. 24).

A recuperação de imagens difere mais de perto da recuperação de textos porque os usuários de bases de dados podem querer pesquisar sobre uma ampla variedade de características, que vão desde as muito exatas (nomes de artistas, títulos de pinturas) até as muito imprecisas (forma, cor, textura). Ao tratar de determinada abordagem, uma base de dados conhecida como MUSEUM, Mehrotra (1997) vê essas características como níveis variáveis de abstração. Os níveis principais são mostrados na figura 98, que Mehrotra explica da seguinte forma:

> Nos níveis mais inferiores estão imagens de bases de dados ou imagens-exemplo. No nível seguinte de descrição, uma imagem é caracterizada em termos de suas propriedades, como cores de último plano/primeiro plano, cores dominantes, histogramas e propriedades de textura. A descrição de imagens em termos de objetos — tais como regiões da imagem, segmentos de limite e contornos — e relações entre eles forma o nível seguinte de abstração. Segue-se o nível de abstração em que as imagens

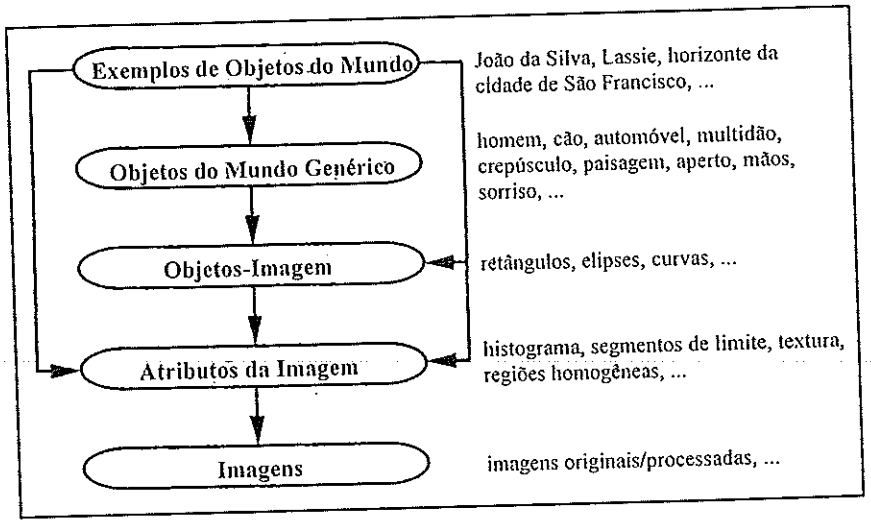


FIGURA 98

Principais níveis de abstração na base de dados de um museu de arte

Apud Mehrotra (1997) com permissão do conselho diretor da University of Illinois

são descritas em termos de objetos genéricos, relações e conceitos, como homem, cão, carro, multidão, horizonte, crepúsculo, nublado, colorido e sorriso. No nível mais elevado de abstração, as imagens são descritas em termos de casos específicos de objetos do mundo genérico. Por exemplo, um homem pode ser descrito como João da Silva, uma cadela pode ser descrita como Lassie, uma imagem pode ser descrita como o horizonte da cidade de São Francisco. As descrições de imagens em qualquer um desses níveis de abstração podem ser multiníveis e ser derivadas das descrições nos níveis inferiores de abstração, ou associadas a elas (p. 61).

As buscas numa base de dados de imagens nos níveis médios de abstração envolvem 'recuperação de imagens baseada em conteúdo'. Continua Mehrotra para caracterizar os requisitos da seguinte forma:

1. *Consultas que não envolvam processamento/análise de imagens* – nestas consultas, não há necessidade de processamento ou análise de imagens da base de dados, e não são apresentadas imagens de consulta. Exemplos: 1) recupere todas as imagens que contenham pelo menos um automóvel em frente de uma casa, 2) recupere fotografias que contenham um homem sorrindo. As descrições simbólicas (extraídas automaticamente e/ou especificadas pelo usuário) relativas às imagens da base de dados são empregadas para selecionar as imagens desejadas. Essas consultas podem ser processadas por meio dos métodos tradicionais.

2. *Consultas que envolvam processamento/análise de imagens* – estas consultas envolvem uma ou várias imagens que são processadas para extrair delas as informações simbólicas desejadas a elas relacionadas. A descrição extraída é comparada com a descrição de imagens da base de dados, a fim de selecionar imagens que satisfaçam às exigências especificadas. Exemplos: 1) recupere todas as imagens que contenham um ou vários objetos similares a determinada imagem de consulta em termos de cor da imagem e características textuais (p. 61-62).

É óbvio que os diferentes níveis de abstração mostrados na figura 98 representam, de cima para baixo, problemas de indexação crescentemente complexos e crescentemente incomuns.

As representações exclusivamente textuais das imagens possuem evidentes limitações. Heller (1974) mostra um exemplo muito radical do registro catalográfico de uma pintura de Picasso (figura 99). O primeiro grupo de elementos do registro representa dados 'exatos' sobre a pintura, mas o segundo grupo, que se refere ao que ali se acha representado, e como é representado, além de ser uma questão de interpretação, oferece uma visão bastante imperfeita de como ela é. Também não inclui outros atributos importantes, principalmente as cores.

Schroeder (1999) descreve como três diferentes 'camadas' de indexação são aplicadas às imagens no General Motors Media Archives: objetos (aquilo que é representado — por exemplo, um caminhão Chevrolet ano 1935), estilo (por exemplo, uma fotografia 'imparcial' *versus* uma fotografia 'atraente' de um veículo) e implicações (por exemplo, ilustra a grande durabilidade do veículo).

É provável que a indexação de imagens por meio de descrições verbais seja ainda mais subjetiva e, portanto, mais incoerente do que a indexação de textos.

Há indícios de que isso seja verdade (Markey, 1984). Isso levou Brown et al. (1996) a sugerir a possível utilidade de uma abordagem 'democrática' da indexação, em que os usuários das imagens sugerem seus próprios termos de indexação, e a fazer experiências com esse método. Vários autores defendem a colaboração dos usuários na indexação de bases de dados de vídeos. Liu e Li (2002), por exemplo, propõem um sistema em que os termos que aparecem nas buscas dos usuários tornar-se-iam termos de indexação relativos aos trechos de vídeo que recuperam (provavelmente apenas os considerados 'relevantes').

É difícil chegar a um acordo sobre a indexação de imagens porque é difícil

Group 1	
TYPE	oil painting
ACCNO	21.37p
TITLE	The Three Musicians
ARTIST	Picasso, Pablo
VENDOR	Gallerie Rosenberg
MADE	1921
LOC	The Museum of Art, New York
SIZE	79 x 87 3/4"
REF	Our-2
REF	70-164877
MEDIUM	oil
MEDIUM	canvas
Group 2	
(5,1)	
FIGURE	musician
POSITION	front
(5,1,1)	
OBJECT	clarinet
COMMENT	left figure
(5,1,2)	
OBJECT	guitar
COMMENT	center figure
(5,1,3)	
OBJECT	accordion
OBJECT	music score
COMMENT	right figure
(5,2)	
FIGURE	dog
POSITION	side

FIGURA 99
Exemplo do registro catalográfico de uma pintura
Apud Heller (1974) com permissão do Strong Museum, Rochester, NY
Esta figura foi reproduzida em Scott (1988)

haver concordância quanto ao que uma imagem realmente mostra. Shatford (1986) faz uma distinção entre *de* que é uma imagem e *do que ela trata*. No primeiro caso, lida-se com coisas concretas (por exemplo, a imagem mostra uma mãe com os filhos), enquanto no segundo caso lida-se mais com abstrações (por exemplo, a imagem mostra miséria, sofrimento, desespero). Em artigo posterior (Layne, 1994), ela identifica vários tipos de 'atributos' na indexação de imagens, embora sugira que disciplinas diferentes podem querer utilizar atributos muito diferentes na indexação de um acervo de imagens. Ela acentua a importância de empregar a indexação para formar grupos úteis de imagens ao invés de pensar somente em imagens tratadas de modo isolado. Krause (1988) trata com certa minúcia do problema da indexação de acervos de imagens. Ele concorda com a distinção entre *de* e *do que* trata, mas adota nomes diferentes, a saber, aspectos *rígidos* [*hard*] e *flexíveis* [*soft*] da imagem.

Svenonius (1994) argumenta que, embora algumas imagens (por exemplo, em textos médicos) destinem-se a transmitir informações, essa não é de fato a finalidade de pinturas e outras formas artísticas. Embora algumas representem pessoas ou objetos que podem ser descritos verbalmente, outras são 'lingüisticamente indeterminadas'.

Markey (1984), Shatford (1986), Svenonius (1994), van der Starre (1995) e Enser (1995), entre outros, referem-se ao trabalho do historiador da arte Panofsky, que sugeriu que uma imagem podia ser analisada do ponto de vista pré-iconográfico, iconográfico e iconológico. Numa experiência de que participaram 18 pessoas, de antecedentes variados, Enser constatou que a mesma imagem seria indexada em todos os três níveis. Por exemplo, uma cena da torre Eiffel receberia termos nos níveis pré-iconográfico (torre, rio, árvore), iconográfico (torre Eiffel, rio Sena) e iconológico (romantismo, férias, emoção). O grande número de termos atribuídos a uma única imagem (18 pessoas atribuíram 101 termos à cena de Paris), argumenta Enser, indica a necessidade de indexação exaustiva.

Orbach (1990) é um dentre vários autores que acentuaram a necessidade de indexar uma coleção de imagens do ponto de vista de determinado grupo de usuários. Em suas próprias palavras:

> A meta da análise temática é capturar a essência de uma imagem ou grupo de imagens — seu conteúdo e temas mais importantes — ao mesmo tempo que permanece alerta para elementos que sabidamente sejam de interesse especial para a clientela do repositório (p. 184).

Para certas exigências, como, por exemplo, recuperação de uma imagem que ilustre uma emoção, a indexação de bases de dados de imagens tem algo em comum com a indexação de obras de ficção, como vimos no capítulo precedente.

Abordagens baseadas no conteúdo

Vários sistemas foram desenvolvidos para permitir a busca de imagens por meio de características de nível baixo, como forma, cor e textura. Na maioria dos

casos, o computador (possivelmente com ajuda humana) extrai das imagens características úteis de nível baixo e recodifica esses dados numa forma simbólica, mais fácil de ser usada em operações posteriores de indexação e recuperação.

O sistema QBIC (Query by Image Content), desenvolvido pela IBM, está sendo empregado em caráter experimental em diversas aplicações (Flickner et al., 1995). Holt e Hartwick (1994), que o utilizaram num contexto de história da arte, descrevem seus recursos da seguinte forma:

> O QBIC oferece várias formas de consultas de imagens. As duas mais gerais são como 'consulta de objeto' ou 'consulta de imagem'. As consultas de objetos recuperam imagens que contêm objetos que coincidem com especificações de consulta, do tipo 'localize formas vermelhas e circulares', enquanto as consultas de imagens buscam a coincidência com características totais de imagens, do tipo 'encontre imagens que possuam principalmente tonalidades de vermelho e azul'. Para efetuar consultas de objetos, estes devem ser identificados em cada cena, normalmente de modo manual, traçando um esboço deles antes da consulta. O processo de esboçar os objetos e em seguida processar atributos ou características de cada objeto e cada imagem como um todo denomina-se classificação de imagens. Há ferramentas básicas de desenho, como retângulo, elipse, polígono, pincel e uma ferramenta de contornos ativos [*snake tool*], que traça o contorno das imagens selecionadas. Uma ferramenta de preenchimento [*fill tool*] acelera o mascaramento de imagens de alto contraste ao traçar automaticamente pixéis de valor similar ao que foi selecionado (p. 82-83).

O QBIC permite a realização de buscas que envolvam cores, texturas e formas, bem como o assunto representado numa pintura. Também permite consulta por exemplo ('encontre outras fotografias como esta'). Holt e Hartwick relatam que buscas sobre formas em pinturas podem enfrentar enormes problemas.

Diversos outros sistemas de recuperação baseados em conteúdo foram desenvolvidos, embora não se tenha clareza sobre quais são 'operacionais' e quais são simplesmente experimentais. Um exemplo característico é o MUSE (Marques e Furht, 2002), um 'protótipo em funcionamento' destinado a suportar pesquisas e consultas por exemplo. Um componente que integra o projeto do MUSE é um mecanismo de retroalimentação de relevância.

As técnicas de reconhecimento e coincidência de formas ainda estão muito aquém da perfeição. E, conforme Picard e Minka (1995) salientam, a análise de formas não resolve todos os problemas de consulta por exemplo — algumas imagens procuradas (um campo, água, multidões, fogo) não possuem uma forma bem-definida, e devem, ao contrário, ser cotejadas pela 'textura'. Eles examinam abordagens de identificação de 'regiões visualmente similares' numa fotografia, empregando características como 'direcionalidade, periodicidade, aleatoriedade, rusticidade, regularidade, aspereza, distribuição da cor, contraste e complexidade'. O sistema experimental que desenvolveram procura imitar o comportamento humano no reconhecimento de cenas visualmente similares. Picard (1996) trata ainda da textura da visão em recuperação de imagens, enquanto Mehrotra e Gary (1995), Mehtre et al. (1997) e Jagadish (1996) tratam do problema do

reconhecimento de formas. Em Ogle e Stonebraker (1995) e Smith e Chang (1997b) encontram-se exemplos de sistemas em que um dos principais componentes é a pesquisa de cores.

Mehrotra (1997) analisa alguns dos problemas presentes na representação e busca de formas de imagens, e Huang et al. (1997) estudam a forma, cor e textura como problemas de indexação e recuperação. O analisador de imagens por eles descrito consegue processar um histograma de cor para uma imagem, bem como uma medida de textura baseada em aspereza, contraste e direcionalidade. A segmentação de imagens é obtida por meio de uma técnica de agrupamento. A posição relativa desses agrupamentos permite buscas nas bases de dados que envolvam cor, textura e características espaciais (por exemplo, 'uma região vermelha acima e à direita de uma grande região azul'). Forsyth et al. (1997) apresentam um amplo e útil panorama sobre o uso de características de cor, textura e geometria na recuperação em grandes bases de dados de imagens.

Mehtre et al. (1998) apresentam um método para o agrupamento de imagens que se baseia numa combinação de características de forma e cor. O grau de coincidência entre qualquer par de imagens pode ser computado e expresso numericamente, permitindo, assim, consultas por exemplo (ou seja, é possível pesquisar imagens similares a outra já selecionada). Alegam um grande sucesso em experiências de recuperação, mas trabalharam com bases de dados muito pequenas (por exemplo, uma delas possuía 500 imagens de logomarcas).

É importante reconhecer, contudo, que a maioria dos usuários de bases de dados de imagens provavelmente não fará buscas sobre aspectos mais abstratos, como cor, forma e textura, embora possam empregá-los para limitar ainda mais uma busca. Huang et al. (1997) assim coloca a questão:

> Em muitas aplicações de sistemas de recuperação de multimídia, os usuários raramente usam características de imagens de nível baixo (isto é, forma, cor, textura) diretamente para consultar a base de dados. Ao contrário, o usuário interage com o sistema mediante conceitos de nível superior (por exemplo, praia, floresta, flores amarelas, crepúsculo) para especificar determinado conteúdo de imagem (p. 115).

Experiências realizadas por McDonald et al. (2001) sugerem que a cor pode ser um critério de classificação e busca bastante útil para o usuário que não tenha em vista determinada imagem.

Diversos sistemas oferecem a possibilidade de consulta por exemplo ou 'recuperação de similaridade'. Kurita e Kato (1993) descrevem várias aplicações experimentais, por exemplo:

1. Ao ser feito o pedido de registro de uma marca, ela pode ser escaneada por um departamento de patentes e cotejada com uma base de dados de marcas existentes.*

* A indexação/recuperação de marcas também é tratada por Wu et al. (1995) e Ravela e Luo (2000), entre outros.

2. Para consultar bases de dados de museus ou museus de arte, o usuário pode esboçar uma imagem (por exemplo, de uma paisagem ou parte de uma paisagem) e o sistema pesquisará as pinturas que mais se pareçam com essa imagem.*

DiLoreto et al. (1995) analisam trabalho que é um tanto similar ao de Kurita e Kato embora em ambiente totalmente diverso. Seu sistema experimental de informação geográfica, 'baseado apenas na representação pictórica de uma consulta', possibilita uma pesquisa que pode envolver a utilização de atributos geométricos, relações topológicas e distâncias.

Nem todos os sistemas baseados em conteúdo estão centrados em imagens em sua totalidade. Continuam sendo realizadas pesquisas sobre métodos para representação e buscas em regiões separadas de uma imagem (ver, por exemplo, Moghaddam et al., 2001). Um livro de autoria de Wang (2001) descreve com detalhes um método 'baseado em regiões' para recuperação de imagens baseada em conteúdos. Esse método é assim descrito:

> Uma imagem, ou parte dela, numa base de dados, é representada por um conjunto de regiões, que corresponde aproximadamente a objetos, que se caracterizam por cor, textura, forma e localização. O sistema classifica as imagens em categorias semânticas, como texturado–não-texturado, censurável–benigno ou gráfico–fotográfico. A categorização melhora a recuperação ao permitir métodos de buscas semanticamente adaptáveis e o estreitamento da faixa de buscas numa base de dados (p. xi-xii).

O método baseado em regiões tem a vantagem de permitir critérios menos estritos para o cotejo de imagens: uma única região numa imagem pode ser comparada com várias regiões em outra imagem. Mesmo que duas imagens não coincidam perfeitamente em sua totalidade, talvez coincidam razoavelmente bem no nível de região.

Jones e Roydhouse (1995) descrevem um curioso sistema, baseado em casos, para indexação e recuperação de dados meteorológicos. Diante de uma situação climática atual, o meteorologista pode pesquisar condições similares em situações passadas. Um mapa das condições atuais (ver figura 100) pode ser usado como uma consulta; o sistema então colocará em ordem de similaridade situações anteriores (ver figura 101). Cada objeto gráfico da consulta (figura 100), como, por exemplo, a localização do centro de pressão e sua magnitude, é convertido numa representação simbólica que é empregada nas buscas na base de dados onde casos anteriores também estão representados simbolicamente.

Os autores descrevem seu método de indexação da seguinte maneira:

> Cada caso representa uma fatia de tempo para a qual se dispõe de dados meteorológicos. Esses dados disponíveis para nós incluem imagens de satélite armazenadas tanto em formato digital quanto em disco laser, um arquivo de documentos e campos numéricos [...] Alguns exemplos de campos numéricos incluem pressão,

* Benois-Pineau et al. (1997) descrevem um método similar no qual as imagens de edifícios podem ser recuperadas pelo cotejo com um 'esboço sintetizado'.

temperatura, umidade relativa, velocidade dos ventos e vorticidade relativa, tudo disponível em 14 níveis diferentes da atmosfera. [O sistema] hoje possui 3,5 anos de dados com intervalos de 12 horas, que constituem uma base com cerca de 2 500 casos. Estamos concentrados atualmente numa região que cobre cerca de uma oitava parte do globo, tendo como centro a Australásia. Prevemos que mais 10 anos de dados logo estarão disponíveis para nós, o que nos permitirá expandir a base de casos para cerca de 10 000 casos. Dentro de alguns anos, as reanálises de dados históricos [...] deverão produzir conjuntos de dados que abrangerão um período desde a Segunda Guerra Mundial até os dias atuais, o que permitirá a construção de uma base de dados com mais de 36 000 casos do passado.

[O sistema] recupera casos por meio do cotejo de consultas feitas pelo usuário com *rótulos de índice* explicitamente representados. As consultas identificam características específicas de nível alto da situação atual que pareçam ser meteorologicamente importantes: por exemplo, sistemas de baixa e alta pressão. Os rótulos de índice são representações das características de nível alto da situação do tempo em cada caso. Tanto quanto possível, [o sistema] extrai essas características automática ou semi-automaticamente dos dados brutos. Atualmente estamos nos concentrando em certas características, como as mínimas e máximas locais, que são fáceis de extrair automaticamente [...] (p. 51).

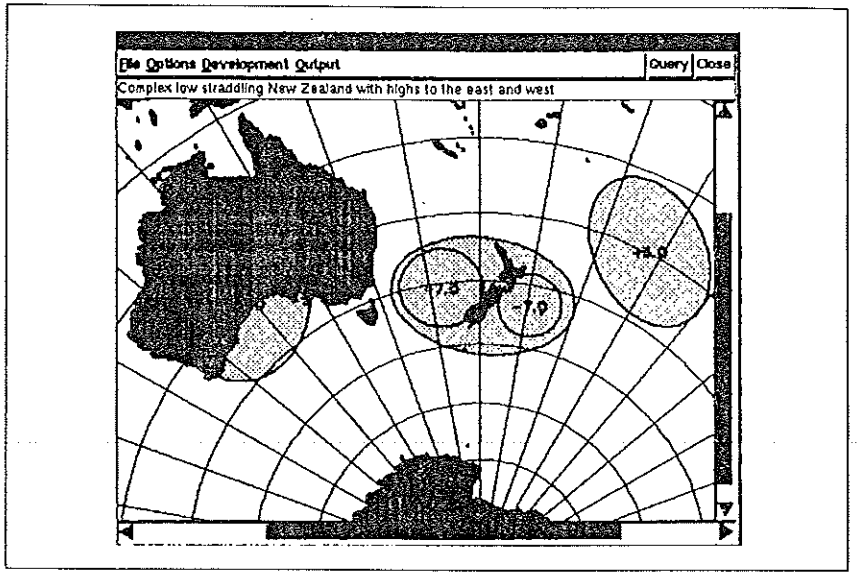


FIGURA 100
Consulta formulada a uma base de dados meteorológicos. A consulta pede um complexo sistema de baixa pressão sobre a Nova Zelândia com sistemas de alta pressão a leste e oeste

Apud Jones & Roydhouse, "Intelligent retrieval of archived meteorological data", *IEEE Expert*, 10 (6), 1995, 50-57. © 1995. IEEE.

Corridoni et al. (1998) descrevem um método de recuperação de pinturas por meio da 'semântica das cores'. Em essência, as pinturas são segmentadas em regiões que possuam diferentes características cromáticas. A base de dados pode então ser consultada para localizar pinturas que apresentem determinadas propriedades cromáticas e características espaciais.

Experiências quanto ao emprego da textura na indexação e recuperação de fotografias aéreas são descritas por Ramsey et al. (1999). O objetivo deles era elaborar um 'tesauro' de texturas (e talvez outras características das fotografias) que os usuários pudessem consultar em linha. Quando o usuário encontrasse uma textura que corresponderia ao elemento procurado (por exemplo, uma pista de aeroporto), ele poderia utilizar o sistema para consultar imagens e encontrar as que apresentassem texturas similares. Alternativamente, seria possível empregar consultas por exemplo; isto é, o usuário solicitaria ao sistema que procurasse fotografias que apresentassem texturas similares às de outra que já tivesse em mãos. Ma e Manjunath (1998) estudam a segmentação e recuperação de fotografias aéreas baseadas em texturas.

Zhu e Chen (2000) chamam atenção para o fato de que um sistema ideal de imagens precisa ter condições de fazer buscas sobre características de nível baixo (como cor, forma e textura) de uma imagem, mesmo que a consulta feita pelo usuário esteja em nível muito mais elevado (por exemplo, encontrar todas as imagens que contenham pomares). Se o usuário selecionar alguma característica (por exemplo, pomar) numa fotografia aérea, o sistema experimental de Zhu e Chen procurará outras imagens que pareçam conter características similares. O sistema emprega apenas textura na comparação de imagens. Sua expectativa é de que resultados muito melhores seriam obtidos se a comparação se baseasse na forma e na cor, bem como na textura.

A indexação de imagens baseada em palavras e feita por seres humanos é cara, e por isso foram feitas várias sugestões sobre como a indexação baseada em conceitos seria efetuada automaticamente, ou, pelo menos, com ajuda do computador. Goodrum et al. (2001) sugerem como características de nível baixo das imagens seriam usadas para agrupar essas imagens de modo a possibilitar a 'herança' de termos. Imaginemos uma coleção de imagens que haja sido indexada com termos atribuídos por seres humanos. A essa base de dados acrescenta-se novo lote de imagens. As técnicas de agrupamento comparam as imagens recém-chegadas (por exemplo, pela forma) com as que já se encontram na base de dados. Se a nova imagem *X* for muito parecida com a imagem antiga *A*, termos *A* seriam atribuídos a *X* também, ou, pelo menos, apresentados como sugestões para indexação de *X*. Propõem também que esse tipo de comparação seja adotado nas atividades de controle de qualidade. Isto é, se a imagem *X* e a imagem *Y* 'fossem parecidas', mas os seres humanos houvessem indexado ambas de modo muito diferente, o sistema criaria um alerta que faria com que fossem examinadas mais cuidadosamente. Por fim, propõem que os usuários de uma base de imagens

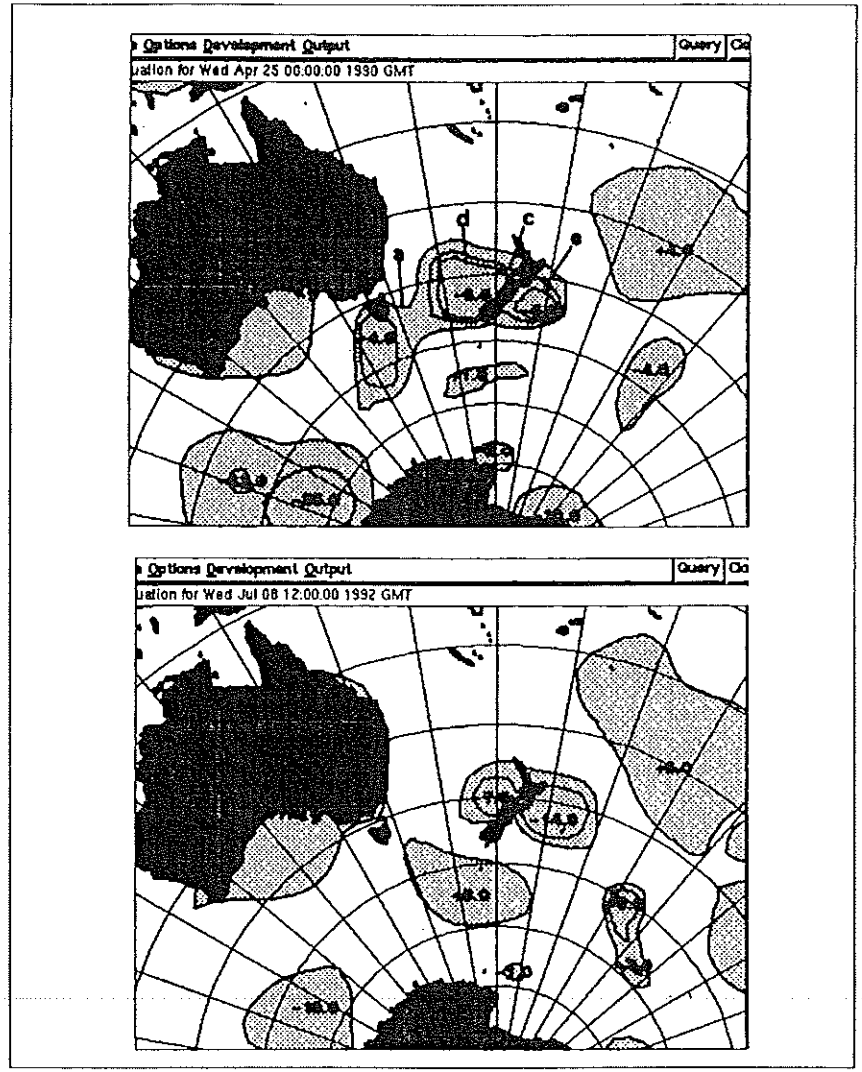


FIGURA 101
Dois mapas meteorológicos recuperados em resposta à consulta da figura 100
Apud Jones & Roydhouse, "Intelligent retrieval of archived meteorological data", *IEEE Expert*, 10 (6), 1995, 50-57. © 1995. IEEE.

sejam solicitados a apresentar uma descrição do uso que pretendem dar a uma imagem (ou grupo) e que essas descrições forneceriam termos que seriam úteis pontos de acesso em futuras recuperações. Patrick et al. (1999) e Frost (2001) também propuseram formas de indexação por 'herança'.

Para imagens presentes num contexto textual (por exemplo, num jornal), talvez seja possível extrair automaticamente partes do texto que expliquem a imagem. Trabalhos nessa linha foram descritos por Srihari (1993, 1995a,b,1997) e Nakamura et al. (1993), entre outros. Estes últimos estudam a integração de informações do texto com informações da imagem (neste caso um diagrama encontrado num manual ou numa enciclopédia). Em seu trabalho, como no de Rajagopalan (1994), o texto é usado para explicar o diagrama. Por exemplo (segundo Rajogopalan), a afirmação 'o disco está rolando caminho abaixo' pode esclarecer muito o que estiver representado num diagrama que é completamente estático. Vários sistemas experimentais 'anotarão' (isto é, indexarão) imagens baseados em palavras-chave que ocorram no texto em volta da imagem. Ver, por exemplo, Liberman et al. (2001).

Srihari volta-se para problemas mais difíceis e sua pesquisa é muito mais complexa, recorrendo ao campo do reconhecimento da fala bem como aos do processamento da linguagem natural e compreensão da imagem. Uma aplicação emprega o texto de legendas para identificar seres humanos retratados em fotografias de jornais. Quando a legenda pode ser usada para identificar um indivíduo, o texto dela é empregado para indexar a imagem automaticamente. No protótipo do sistema denominado Show & Tell (Srihari, 1997), um analista humano vê a imagem de uma paisagem numa estação de trabalho e a descreve (indexa) mediante uma combinação de entrada de dados com o mouse (apontamento) e linguagem falada. Um sistema de reconhecimento da fala transcreve a entrada e a sincroniza com a entrada de dados feita pelo mouse. Esse tipo de 'videoanotação' foi expandido para um sistema destinado à anotação de quadros de vídeo com especial referência à indexação e buscas em vídeos em aplicações de inteligência militar.

Carrick e Watters (1997) apresentam um método para problema afim: o reconhecimento automático de associações entre diferentes mídias, como no reconhecimento de que determinada fotografia se relaciona com determinada notícia.

Parece provável que alguns usos das bases de dados de imagens serão tão imprecisos que somente permitirão os métodos de pesquisas aleatórias ou buscas iterativas. Um exemplo óbvio é a busca de um rosto do qual se conhecem ou são lembrados apenas os traços gerais.* Jain (1997) examina este problema e o método de busca iterativa para resolvê-lo (chama-o de 'consultas incrementais'):

> O usuário que estiver à procura de certas informações, por exemplo, acerca de uma pessoa de quem tem uma vaga lembrança, especifica coisas importantes que ele recorda sobre a pessoa [ver figura 102]. Esta especificação talvez diga que ela tem olhos grandes, boca grande, cabelo longo e testa pequena. Com base nessas informações, recuperam-se fotografias de pessoas que nelas se enquadrem. O usuário

* Vários métodos de reconhecimento de fotografias de rostos são analisados na literatura. Por exemplo, Rickman e Stonham (1991) propõem um método baseado em rede neural. O problema também é abordado por Wu et al. (1995), Pentland (1997), Li et al. (1997), Hafed e Levine (2001) e Fleuret e Geman (2001).

poderá, então, selecionar a que mais se aproxime de sua consulta e modificar a consulta seja especificando características seja empregando na fotografia ferramentas de edição gráfica e de imagens. Isso refina a imagem de consulta, que é então enviada ao sistema para que forneça novos candidatos à satisfação da consulta. Assim, a consulta é formulada de modo gradativo, começando com a idéia vaga original. Esse processo será concluído quando o usuário considerar-se satisfeito (p. 71).

Price et al. (1992) avaliam um método de busca iterativa (retroalimentação de relevância) para a recuperação de imagens, mas baseado em descrições textuais das imagens ao invés de buscas de coincidência de padrões das próprias imagens. Gudivada et al. (1996) tratam da retroalimentação de relevância em relação à última situação. Ciocca e Schettini (1999) apresentam um método de modificação de busca automática baseado nas características de nível baixo das imagens selecionadas pelo usuário como úteis e das selecionadas como não-úteis.

Rowe e colegas da U.S. Naval Postgraduate School formam um grupo de pesquisa que se concentrou na indexação de fotografias e outras imagens. Seu método emprega uma combinação de texto (legendas de figuras) e processamento de imagem em nível de pixéis. Um método de rede neural é usado para classificação de regiões aplicada a fotografias, e processos de análise [*parsing*] automática são aplicados às legendas. Seu trabalho, que se concentra em dados multimídias em sistemas de armamentos, inclui a indexação de fotografias que fazem parte de páginas da Rede (Rowe e Guglielmo, 1993; Rowe, 1994, 1996; Rowe e Frew, 1996, 1997; Guglielmo e Rowe, 1996).

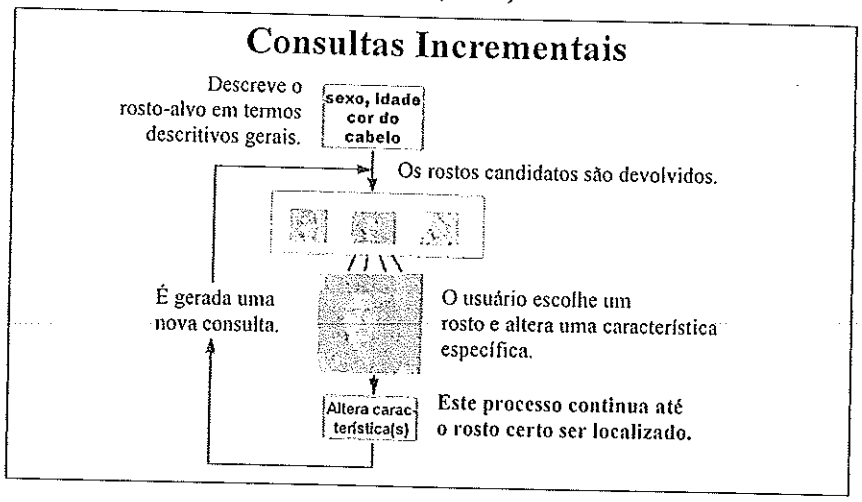


FIGURA 102
Consulta incremental numa base de dados de imagens
Apud Jain (1997) com permissão do conselho diretor da University of Illinois

Gauch et al. (1999) descrevem um sistema — VISION — que atribuirá trechos

de vídeo a categorias baseadas nos termos que ocorrem em suas legendas. O esquema de classificação adotado contém cerca de 2 000 categorias. Os vídeos que chegam podem ser cotejados com perfis de interesses dos usuários mediante este conjunto de categorias.

Vailaya et al. (2001) desenvolveram procedimentos para colocação de imagens em categorias baseadas em suas características de nível baixo. Os experimentos que descrevem empregam uma base de dados que contém quase 7 000 fotografias de viagens de férias:

> Especificamente, estudamos a classificação hierárquica de imagens de viagens de férias; no nível mais alto, elas são classificadas como internas ou externas; as externas são ainda classificadas como urbanas ou naturais; por fim, um subconjunto de imagens naturais é classificado nas classes de crepúsculo, floresta e montanha (p. 117).

A classificação baseia-se na distribuição da cor e características da forma.

Vários grupos de pesquisa vêm estudando métodos para indexação de coleções de pinturas e outros objetos de arte. Por exemplo, Ozaki et al. (1996) descrevem uma abordagem que incorpora informações sobre o que é representado (por exemplo, orientação espacial) bem como sobre fatores estéticos, como cor e estilo.

Encontram-se na literatura trabalhos que lidam com problemas mais complexos da recuperação de imagens. Por exemplo, Cromp e Dorfman (1992) examinam um método para lidar com dados de sensoriamento remoto obtidos por satélites em órbita, e Gudivada e Raghavan (1995) identificam situações complexas, em matéria de recuperação, relativas a certos tipos de bases de imagens, inclusive a representação e recuperação de imagens tridimensionais ('recuperação por volume') e 'recuperação por movimento' (por exemplo, encontrar uma imagem que mostra determinada ação).

Geisler et al. (2001) descrevem trabalho em curso na University of North Carolina visando ao desenvolvimento de uma coleção digital de vídeos (o Open Video Project) que pode ser empregado como bancada de provas para investigações sobre pesquisas, recuperação e uso de segmentos de vídeo digital.

Em livro de Wu et al. (2000) encontra-se uma análise exaustiva (mas altamente técnica) de vários aspectos da recuperação baseada em conteúdos.

Embora alguns pesquisadores da área da recuperação baseada em conteúdos sejam bastante ousados em suas pretensões, outros são bem modestos. Por exemplo, Wang (2001), escrevendo sobre seu trabalho com recuperação de imagens baseada em conteúdos, na Stanford University, na década de 1990, admite:

> Na época, a mim parecia razoável que haveria de descobrir a solução para o problema da recuperação de imagens no curso do projeto. A experiência mostrou, com certeza, que ainda estamos longe de resolver esse problema básico (p. xi).

Imagens na Rede Mundial

É natural que atualmente se esteja dando muita atenção a maneiras de melho-

rar a recuperação de imagens na Rede Mundial. A maioria dos mais importantes mecanismos de buscas realmente oferece recurso que restringe uma busca verbal à recuperação de imagens. No entanto, as pesquisas atuais tratam de processos mais complexos, como a identificação automática de imagens. Um livro de Chang et al. (2001) contém uma descrição mais atualizada dos métodos de buscas de multimídia na Rede.

Iyengar (2001) organizou uma série de artigos sobre acesso a imagens na Rede. Chen et al. (2001) descrevem seu método de extração de informação textual de imagens na Rede (a partir de Localizadores Universais de Recursos [URLs], títulos, textos ao redor de uma imagem). Esses atributos semânticos de nível alto podem então ser combinados com atributos de nível baixo. Liu et al. (2001) descrevem mais detidamente a extração de texto. Outro método que combina características de nível baixo e nível alto é descrito por Wu et al. (2001).

Rowe e Frew (1998) descrevem métodos desenvolvidos para identificação automática de fotografias na Rede Mundial. As fotografias podem ser identificadas mediante uma combinação de características, inclusive forma, dimensões, quantidade de cores e referências do texto. A segunda etapa dessa pesquisa é a identificação automática de legendas para essas fotografias. Essa não é uma tarefa simples, porque, na página da Rede, as legendas podem estar separadas das fotografias, e às vezes inseridas em texto maior. A localização de legendas é feita por meio do emprego de "chaves multimodais que incluem as palavras específicas utilizadas, a sintaxe, o leiaute circundante da página da Rede, e a aparência geral da imagem associada". Os autores reivindicam "um surpreendente grau de sucesso" de procedimentos que evitam o processamento da imagem completa e processamento total da linguagem natural.

Vários grupos de pesquisas vêm trabalhando sobre reconhecimento de fotografias de pessoas na Rede. Os métodos podem basear-se no reconhecimento facial e ocorrência do nome no texto, ou uma combinação de ambos (ver, por exemplo, Aslandogan e Yu, 2000).

Agnew et al. (1997) descrevem um método experimental de consulta por exemplo para busca de imagens na Rede Mundial. O sistema localizará as imagens, fará sua indexação (por cor, tamanho e outros atributos) e armazenará os índices num servidor. Smith e Chang (1997a) estudam outro método de indexação de imagens na Rede, que emprega tanto atributos textuais quanto visuais.

Resumos de imagens

A preparação de um resumo, ou outro tipo de sucedâneo, de uma imagem apresenta problemas especiais, principalmente no caso de imagens em movimento, como os programas de televisão. Basicamente, são possíveis dois tipos de resumos: uma descrição verbal do vídeo (um resumo bastante convencional) ou um resumo que seja ele próprio uma imagem. Embora seja possível elaborar resumos textuais que sintetizem as ações dos filmes ou transmissões de televisão

(como o demonstra o catálogo do American Film Institute), talvez seja melhor, em certos casos, dispor de um resumo visual de cenas do próprio filme. Geisler et al. (2001) salientam que os resumos de vídeo podem ter o formato de imagens estáticas ou em movimento. Assim se referem aos resumos de imagens estáticas:

> *Slide shows*, *storyboards* e tiras de filmes [*filmstrips*] são exemplos desse tipo [...] Normalmente as pessoas extraem os quadros-chave [*keyframes*] de cada tomada para representá-la e em seguida arranjam todos os quadros-chave ou um subconjunto deles para formar o resumo. Os métodos de seleção de quadros-chave e agrupamento ou montagem deles variam em diferentes projetos (p. 68).

E acrescentam:

> Um resumo de imagens em movimento é em si mesmo um vídeo curto e pode oferecer aos usuários informações ricas e animadas. O exemplo mais reconhecível é o *trailer* de filmes [...] O Movie Content Analysis Project [...] seleciona alguns trechos de um filme e em seguida os monta no resumo final. Resumos de imagens em movimento incorporam tanto informações de áudio quanto visuais de uma fonte mais longa e podem ser considerados uma pré-visualização curta de um vídeo longo (p. 68).

Geisler et al. alegam que ainda é preciso pesquisar bastante sobre como as pessoas interagem com os videorresumos.

Ding et al. (1999) compararam três tipos de videorresumos — quadro-chave, verbal (palavra-chave/frase) e uma combinação de ambos — com base na compreensão verbal (a capacidade de a pessoa apreender a idéia principal de um trecho de vídeo a partir do resumo) e 'a essência visual' [*visual gisting*]. No último, foram mostradas imagens aos sujeitos do teste, algumas extraídas do vídeo de origem e outras não, tendo sido solicitado a eles que selecionassem as que pertenciam ao vídeo de origem. Trabalho relacionado a este foi relatado por Tse et al. (1999), que estudaram os efeitos de diferentes visualizações de quadros-chave no desempenho do usuário em tarefas de localização de informação. Os usuários consideraram a visualização estática (*storyboard*) mais fácil de utilizar do que a visualização dinâmica (*slide show*), embora não hajam sido encontradas diferenças no desempenho da tarefa.

Goodrum (2001) comparou quatro tipos de sucedâneos de vídeos (título, palavras-chave, quadros estáticos e quadros-chave) cotejando as decisões de semelhança para cada sucedâneo com as decisões de semelhança para os vídeos representados, na hipótese de que o melhor sucedâneo é aquele cujo 'mapa' de semelhança mais se aproxima do mapa dos próprios vídeos. Houve maior concordância dos sucedâneos baseados em imagens do que dos que se baseavam em textos. Goodrum, no entanto, conclui que há necessidade de ambos:

> Parece claro que, apesar de os sucedâneos baseados em imagens terem alcançado, em geral, melhor desempenho, os sistemas de recuperação de vídeo não devem excluir as representações textuais. Cada tipo de sucedâneo tem uma contribuição exclusiva a dar à percepção, pelo usuário, do conteúdo informacional, e deve ser incluído como parte de um sistema completo de recuperação de informações visuais (p. 11).

Lienhart et al. (1997) descrevem, da seguinte maneira, seu método de elaboração de resumos de vídeos:

> O algoritmo de resumos que desenvolvemos pode ser subdividido em três passos consecutivos [...] No primeiro passo, segmentação e análise do vídeo, o vídeo de entrada é segmentado em suas tomadas e cenas. Ao mesmo tempo, identificam-se seqüências de quadros com eventos especiais, como um texto que aparece na seqüência de créditos, tomadas em *close-up* dos atores principais, explosões e tiros. No segundo passo, seleção dos trechos, selecionam-se trechos do vídeo para inclusão no resumo. O terceiro, montagem dos trechos, monta-os em suas seqüências finais e produz o leiaute de apresentação; este passo envolve a definição da ordem dos trechos do vídeo, o tipo de transição entre eles e outras decisões de edição (p. 56).

Foram desenvolvidos sistemas para selecionar automaticamente quadros-chave de vídeos e incorporá-los numa interface de busca ou consulta em bases de dados de vídeos (ver, por exemplo, Girgensohn et al., 2001). Isso equivale mais ou menos a colocar resumos numa interface que facilite as buscas e a consulta de textos.

Vários grupos de pesquisa estão trabalhando no desenvolvimento de resumos de seqüências de vídeo que sejam eficazes e 'dinâmicos'. Exemplo disso é o trabalho de Nam e Tewfik (2002), que critica os resumos de vídeos que dependem de arranjos estáticos de quadros-chave apresentados em tela única. Tais resumos não 'preservam a natureza dinâmica do passar do tempo' do conteúdo do vídeo. Propõem um método de sumarização* que produza um resumo dinâmico do vídeo. Isso seria conseguido por meio de um processo de amostragem que selecionaria segmentos do filme com base na quantidade de 'atividade' representada. O resumo do vídeo 'apresenta o conteúdo essencial dos dados presentes no vídeo por meio de uma rápida reprodução seqüencial'.

Atributos da imagem

O grande aumento do interesse pela indexação e recuperação de imagens suscitou inúmeros estudos sobre como as pessoas vêem as imagens ou reagem a elas. Esses estudos destinam-se a descobrir quais os tipos de abordagem que as pessoas necessitarão na recuperação de imagens e quais os tipos de termos que serão úteis para descrever e indexar imagens.

Jörgensen (1998) solicitou a 48 mestrandos que 'descrevessem tarefas', a fim de identificar atributos de imagens que seriam úteis na indexação e recuperação. Foram mostradas aos estudantes as mesmas seis imagens e lhes foi solicitado que redigissem uma 'descrição simples' de cada uma, bem como uma consulta para a qual a imagem seria uma resposta coincidente. Jörgensen relata que embora os atributos que ocorriam com mais freqüência (termos para objetos

* No sentido dicionarizado de *resumir*. *Sumarização*, no sentido de elaboração automática de resumos, é empregada por especialistas lusófonos da área da lingüística computacional. (N.T.)

e pessoas representadas, partes do corpo, roupas, cor e localização) fossem previsíveis e coerentes com estudos anteriores, os termos que descreviam a 'história' na fotografia foram usados muito mais do que seria natural.

Heidorn (1999) estudou a descrição em linguagem natural de objetos (neste caso, fotografias de árvores floridas) criada por pessoas "que tentavam descrever objetos de forma suficientemente minuciosa para que o ouvinte reconhecesse o objeto num conjunto de objetos similares". Ele descobriu que os participantes faziam grande uso de analogias em suas descrições (por exemplo, uma planta que parecia uma borboleta).

Goodrum e Spink (1999) examinaram mais de um milhão de consultas por imagens feitas por 211 000 usuários de um único mecanismo de busca na Rede, o EXCITE. Constataram que, em média, havia 3,74 termos por consulta e que a grande maioria de termos empregados eram exclusivos, com mais da metade ocorrendo apenas uma vez.

Frost (2001) estudou usuários que faziam buscas numa base de imagens, quando estavam disponíveis tanto as opções visuais quanto verbais. Os sujeitos da pesquisa eram estudantes, funcionários e membros do corpo docente de uma universidade. A base de dados incluía imagens relativas à Terra e às ciências espaciais. Um dos principais objetivos da pesquisa era determinar se os usuários possuíam uma imagem mental daquilo que estavam procurando e se a imagem recuperada coincidia com a imagem mental. Com base em resultados preliminares, ela concluiu que apenas a recuperação baseada em conteúdo não era suficientemente boa para os usuários generalistas, enquanto apenas a recuperação baseada em conceitos exigia mão-de-obra intensiva. Os custos de um sistema de recuperação de imagens seriam reduzidos se somente parte da coleção fosse indexada. Os usuários encontrariam uma imagem satisfatória nessa parte e a utilizariam para uma busca visual na parte maior da coleção.

Burke (2001) relata estudos sobre classificação de fotografias. Ela empregou a 'teoria do construto pessoal' (uma técnica importada do campo da psicoterapia) em seus exercícios de classificação, e encontrou "um alto nível de coerência entre os construtos pessoais que os participantes empregaram para distinguir as fotografias umas das outras".

O'Connor et al. (1999) realizaram experiências em que estudantes eram solicitados a observar imagens selecionadas e registrar as reações que sentiam diante dessas imagens. A hipótese subjacente era que essas reações seriam uma fonte útil de descritores para a organização de uma coleção de imagens, de modo a facilitar a recuperação futura (isto é, indexação centrada no usuário). Foi pedido aos estudantes que redigissem legendas e também anotassem palavras ou frases que descrevessem o que a imagem continha e o que sentiram diante das imagens. Um resultado observado pertinente à indexação de imagens foi a ocorrência, não incomum, de antonímia: uma frase empregada por um estudante para descrever uma imagem era quase diametralmente oposta à empregada por

outro estudante (por exemplo, um pato 'que apenas nadava', na visão de um estudante, era visto por outro como se estivesse 'numa missão'). Embora o uso de termos de 'reação' talvez seja útil na indexação e recuperação de imagens, pelo menos como suplemento a termos mais convencionais, descritivos (como 'pato' e 'lago'), é claro que teriam de ser fornecidos por uma amostra representativa de observadores, a fim de captar diferentes interpretações e pontos de vista.

Com base em análise de quais os tipos de termos que os usuários de uma amostra empregariam ao observar imagens selecionadas, Greisdorf e O'Connor (2002) concluem que "termos de consulta de base afetiva/emocional parecem ser uma categoria descritiva importante na recuperação de imagens". É difícil entender a lógica de tal afirmativa. Os termos afetivos/emocionais (entre os exemplos citados estão 'bonito', 'sempre jovem', 'feliz', 'forte', 'melancolia') devem certamente corresponder a reações totalmente dependentes do momento temporal. Isto é, se a pessoa *A* julga que determinada imagem sugere a idéia de 'forte', haverá alguma probabilidade de sua reação ser a mesma depois de um ano? Os autores não procuraram estudar a estabilidade desse tipo de reação ou mesmo a coerência da reação entre um grupo grande de pessoas, de modo que carecem por completo de base que sustente sua conclusão. Ademais, é muito difícil acreditar na probabilidade de usuários de uma base de imagens fazerem grande uso desses termos em buscas reais. 'Estou à procura de uma fotografia de árvores que sugira a idéia de 'forte'' parece ser algo completamente implausível.

Choi e Rasmussen (2002) recorreram a membros dos corpos docente e discente de pós-graduação de departamentos de história de duas universidades em seu estudo sobre critérios para determinar a relevância de uma imagem para uma necessidade de informação. As consultas situavam-se no campo da história norte-americana. Naturalmente, a 'topicalidade' (isto é, a imagem guarda relação com a tarefa do usuário) foi o critério mais importante nos julgamentos de relevância, embora outros critérios, como qualidade e clareza da imagem, também fossem importantes. Como também foi notado em muitos estudos nessa área, as conclusões a que chegaram os autores são relativamente triviais. Por exemplo, concluem que é mais provável que os usuários julguem a relevância das imagens a partir das próprias imagens do que a partir das descrições textuais dessas imagens:

> Em primeiro lugar, os sistemas de recuperação devem permitir aos usuários compulsar e comparar um conjunto de imagens recuperadas, pois a visualização das imagens torna mais fácil os julgamentos de relevância (p. 715).

Também concluem que a retroalimentação de relevância pode ser mais importante na recuperação de imagens do que na de textos. Embora isso possa ser verdadeiro, não foram coletados dados que fundamentem tal conclusão. Conforme documentado antes neste capítulo, vários sistemas experimentais de fato incorporam a retroalimentação de relevância.

Chen (2001a,b) estudou as consultas elaboradas por estudantes para localizar imagens necessárias à preparação de trabalhos finais de curso no campo da

história da arte. Chen faz questão de nos dizer que os estudantes "raramente usaram os conceitos de cor, forma e textura em suas consultas", implicando, talvez, que essas características de nível baixo seriam pouco pertinentes à recuperação de imagens em história da arte. Porém os temas atribuídos aos estudantes (por exemplo, o papel de Veneza na história da arte) não eram do tipo que provavelmente exigiria, na recuperação, uma abordagem baseada em conteúdo. Além do mais, nenhum sistema de recuperação de imagens foi realmente utilizado no estudo, e as fontes de imagens disponíveis para os estudantes (ferramentas impressas e sítios da Rede) não foram projetadas para permitir buscas baseadas em conteúdo, de modo que ficamos a imaginar por que essa conclusão viria a merecer qualquer destaque.

Turner (1995) comparou os termos que os usuários selecionaram para aplicar a imagens (neste caso tomadas de filmes cinematográficos) com termos já associados com as imagens na indexação ou em descrições escritas da tomada. Ele encontrou alto nível de concordância. Keister (1994) contribuiu com uma proveitosa análise dos tipos de consultas feitas a uma base de dados de imagens, neste caso estampas e fotografias médicas de interesse histórico, e Sutcliffe et al. (1997) estudaram estratégias de busca de informações adotadas por usuários de bases de dados de multimídia. Hastings (1995a,b,c) estudou os tipos de pontos de acesso de que precisam os historiadores da arte. Depois de observar uma pequena coleção de imagens de pinturas, os historiadores foram entrevistados com a finalidade de determinar, entre outras coisas, qual o tipo de ponto de acesso que lhes seria útil. Ornager (1997) estudou as necessidades de jornalistas no uso de um arquivo de imagens de jornal.

Jörgensen (1996) constatou que sujeitos solicitados a descrever imagens tinham maior probabilidade de selecionar atributos 'perpétuos' (isto é, características bastante exatas, tais como objetos representados e sua cor) ao invés de 'interpretativos' (por exemplo, estilo artístico ou 'clima' de uma pintura) ou atributos 'criativos' (isto é, reação pessoal à pintura, como julgá-la feia ou perturbadora). No entanto, quando instados por meio de um 'gabarito' que apresentava uma série de atributos de todos os tipos, os sujeitos mostravam maior diversidade dos atributos selecionados. Ela conclui, a partir disso, que a indexação eficaz de imagens requer o emprego de uma ampla gama de atributos: perceptuais, interpretativos e reativos.

Com base em conceitos ou em conteúdo?

Layne (2002) é bastante crítico dos métodos completamente automáticos de indexação de imagens:

> Quem ou o que faz a análise do assunto numa obra de arte? Há alguns anos surgiu um forte interesse pela análise informatizada de imagens, e foram encetadas várias tentativas de aplicar técnicas de reconhecimento de padrões e métodos iterativos à identificação e recuperação de imagens relevantes. Até agora, nenhum desses esforços

> teve êxito na recuperação de imagens de grupos heterogêneos ou na identificação de objetos, como cavalos, que podem ser representados em diversas poses, a partir de muitos ângulos diferentes e sob variadas condições de iluminação. Os sistemas informatizados são mais bem-sucedidos na análise de conjuntos homogêneos de imagens e na seleção de imagens com base exclusivamente na cor, composição e textura. Tais elementos são relativamente fáceis de codificar e, portanto, de identificação relativamente fácil pelo computador. Foi aparentemente com grande esforço que alguns sistemas tiveram algum sucesso na identificação de tipos de imagens, como paisagens, que tendem a apresentar certas características comuns de cor e composição. Mas é seguro dizer que a recuperação de imagens baseada em conteúdo — ou seja, informatizada — ainda está longe de vir a ser útil, mesmo remotamente, para historiadores e pesquisadores de arte. [...] Parece que, por ora, o ideal seria deixar o homem fazer o que sabe fazer e o computador fazer o que sabe fazer. Em outras palavras, que o homem identifique os assuntos de uma imagem artística e que o computador identifique cor, forma e composição. Por exemplo, se o indexador humano identificasse os assuntos de imagens de arte, o computador analisaria, se necessário, um grande conjunto recuperado de imagens do mesmo assunto (por exemplo, 'catedrais', 'dança', 'sarcófagos') em busca de semelhanças de forma, cor ou composição (p. 14-15).

A maioria dos autores parece concordar com que a recuperação eficaz de imagens exige tanto métodos baseados em conceitos quanto baseados em conteúdo, aspecto esclarecido por Ornager (1994):

> Embora a idéia de dar entrada a uma imagem-consulta tenha muitos argumentos que a recomendam, essas imagens-consulta nem sempre substituem a força descritiva das palavras, que podem ser melhores para alguns conceitos abstratos. É difícil perceber como seria possível criar uma imagem-consulta que representasse, por exemplo, 'despovoamento de pequenas aldeias norueguesas' ou 'ciúme' (p. 214).

Cawkell (1993) focalizou o mesmo tópico:

> As imagens-consulta substituirão as imagens descritas com palavras à medida que as técnicas forem sendo aperfeiçoadas, embora nem todos os conceitos possam ser assim consultados de modo melhor. Consultas do tipo 'Quais as pinturas que mostram senhoras portando medalhas?', ou 'Há alguma pintura do século XVII onde apareçam animais de estimação?' seriam bem-sucedidas. Mas conceitos abstratos talvez sejam mais bem expressos por meio de palavras usadas como descritores (p. 409).

Turner (1990) salientou que, mesmo que se possa ter acesso muito rápido a uma imagem (neste caso quadros de filmes cinematográficos), isso não exclui a necessidade de acesso a uma descrição textual:

> Além do mais, muitas vezes o texto funciona como um guia da imagem. Em muitos casos, consultar uma sinopse ajuda o observador a interpretar a imagem; por exemplo, talvez seja útil saber que o trem que a pessoa observa é o Expresso do Oriente, ou que a favela que aparece na tela fica bem na periferia de Quito. Em outras palavras, o texto de uma sinopse visual pode proporcionar informações úteis que não estão disponíveis na imagem. Assim, embora seja certamente conveniente ter acesso instantâneo à imagem, isso não dispensaria a necessidade de uma sinopse textual (p. 7).

No mesmo diapasão, Green e Klasén (1993) descrevem as experiências da Sveriges Television [Televisão Sueca] com a indexação de programas de televisão, por meio, exclusivamente, de descrições textuais. Todas as cenas com mais de dez segundos de extensão são descritas com anotações em texto livre, como, por exemplo:

> Rua de feira, apinhada de gente. Barraca de feira, laranjas, maçãs, uvas, pêssegos. Uma caixa de batatas cai no chão. Batatas rolam nas pedras do calçamento. Moça leva as mãos ao rosto.

Trant (1995) assevera que "a descrição textual permanece sendo a chave da recuperação de imagens", acentua a necessidade de uma norma sobre como descrever imagens em bases de dados de imagens e menciona trabalho realizado visando ao desenvolvimento dessa norma.

Mostafa e Dillon (1996) testaram uma interface de um sistema de recuperação de imagens que possuía recursos tanto para buscas visuais quanto verbais. Concluíram que era provável que seus sujeitos (18 estudantes) utilizassem mais o método verbal do que o visual, e sua sugestão era de que isso poderia ser devido basicamente à falta de familiaridade com o método visual.

Ogle e Stonebraker (1995), ao analisar sua experiência com um grande sistema de recuperação de imagens na University of California, Berkeley, reconhecem que "o melhor resultado na recuperação é obtido quando critérios de buscas baseados em textos são combinados com critérios baseados em conteúdo".

O texto ainda é essencial mesmo para as mais avançadas aplicações de recuperação de multimídia. Por exemplo, Hauptmann e Witbrock (1997) utilizam transcrições da parte de áudio dos noticiários de televisão como um meio para recuperação de segmentos de notícias, para atender a pedidos (é utilizada a tecnologia de reconhecimento da fala para criar as transcrições e também para possibilitar consultas faladas), e Mani et al. (1997), em pesquisa assemelhada, utilizam texto de legendas fechadas na recuperação de vídeo de noticiários. O texto de legendas fechadas é usado de forma similar por Takeshita et al. (1997).

Mesmo o sistema experimental de recuperação em arte analisado por Kurita e Kato (1993) não depende inteiramente do exemplo visual para fins de busca. Uma alternativa é a 'consulta por descrição subjetiva', que envolve a indexação das pinturas com adjetivos que representem 'impressões' do observador (por exemplo, 'quente', 'brilhante', 'japonizado'). Do mesmo modo, DiLoreto et al. (1995) incorporam recursos de consulta tanto visuais quanto descritivos em seu sistema de recuperação geográfica.

Cawkell (1994) foi um dos que focalizaram o problema da recuperação baseada exclusivamente em conteúdo:

> Quanto mais complexas as imagens mais difícil fica para o usuário produzir um exemplo visual utilizável, e mais difícil se torna efetuar o cotejo de padrões. Talvez seja preciso cotejar padrões tridimensionais; isso aumenta as dificuldades. A ordem

de dificuldade cresce ainda mais se o usuário estiver interessado em recuperar imagens que contenham determinado objeto dentro de uma imagem.

Por exemplo, se o usuário quiser recuperar 'todas as imagens onde houver um automóvel', não seria muito difícil representar um carro com o auxílio dos programas atualmente disponíveis que incorporam arquivos de *clip art* (que contêm uma grande seleção de objetos desenhados) e programas do tipo 'ferramenta de desenho'. Quando a 'consulta-imagem' é submetida à base de dados para efetuar a comparação, deverá ser possível recuperar um carro que esteja em qualquer imagem, independentemente de como esteja representado e posicionado — tarefa que não é impossível, mas atualmente lenta, que exige o uso, intensivo e caro, de computadores (p. 129).

É claro que há grande diversidade nas aplicações de recuperação de imagens e é provável que nem todas tenham muito a ganhar com alguma forma de indexação baseada em conteúdo. Um estudo sobre demanda de fotografias por jornalistas, realizado por Markkula e Sormunen (2000), encontrou muito pouca necessidade de um método de recuperação baseado em conteúdo, embora os pesquisadores hajam tentado inventar alguns usos possíveis. Os jornalistas, na realidade, externavam necessidades muito simples (por exemplo, fotografias de objetos ou pessoas cujo nome era conhecido), mas não está claro até onde isso teria sido influenciado por limitações conhecidas na indexação do arquivo fotográfico.

Wang (2001) proporciona um resumo muito útil dos tipos de consultas com que devem lidar os sistemas baseados em conteúdo:

- Consulta tipo histograma: encontrar imagens com 50% de vermelho e 20% de amarelo ...
- Consulta tipo leiaute: encontrar imagens que tenham na parte superior um objeto azul e na parte inferior um objeto verde ...
- Consulta tipo forma: encontrar imagens que tenham três estrelas triangulares amarelas dispostas em anel ...
- Consulta tipo esboço desenhado à mão: encontrar imagens que pareçam com determinado desenho ...
- Consulta por exemplo: encontrar imagens que pareçam determinada imagem ... (p. 19)

No entanto, ele salienta a seguir que a maioria dos usuários de imagens estará mais interessada em buscas em 'semântica de nível alto':

- *Objeto*: contém uma lesão
- *Relação do objeto*: contém uma lesão perto do líquido cerebrospinal
- *Clima*: uma imagem feliz
- *Tempo/Lugar*: noite em Yosemite
- (p. 19-20)

Enser (2000) argumenta que as buscas baseadas em conceitos continuarão a predominar sobre as exigências dos usuários em coleções de arquivos de imagens, mas que demandas menos tradicionais por informações visuais (por exemplo, coincidência de impressões digitais e logomarcas, reconhecimento facial, classificação baseada em textura de imagens geológicas) exigem um método baseado

em conteúdo. O ideal é um sistema híbrido — em que uma busca verbal seja usada para recuperar imagens relevantes e estas possam então ser usadas para procurar imagens semelhantes com base em características de conteúdo.

Em conclusão, o método ideal de recuperação de imagens talvez seja aquele que combine acesso convencional por meio de texto (termos de indexação ou narrativa descritiva) com o cotejo de imagens. Assim, uma busca com palavras (batalha, ataque, luta) recuperaria uma imagem de determinado tipo de cena e esta, por sua vez, poderia ser usada como insumo para localizar outras iguais. Uma abordagem possível é um tesauro visual — um tesauro que armazene imagens representativas junto com rótulos verbais (Seloff, 1990) ou possivelmente sem os rótulos verbais. Para uma análise das vantagens e características dos tesauros visuais, nas buscas em bases de imagens, ver Hogan et al. (1991).

Chu (2001), com base em análise bibliométrica da literatura, conclui que não se verificou suficiente interação entre os que trabalham com a abordagem baseada em conteúdo e os que trabalham com a baseada em conceito, embora a situação possa estar melhorando.

Metadados e vocabulários de indexação

Um livro organizado por Baca (2002) trata de metadados e vocabulários controlados na descrição de imagens de arte. As ferramentas mostradas incluem Categories for the Description of Works of Art (Harpring, 2002) e ICONCLASS (Hourihane, 2002), sendo este um esquema de classificação, com notação, para a descrição de pessoas, objetos e atividades representados em obras de arte.

Esquemas de metadados aplicáveis a imagens digitais são revistos por Greenberg (2001).

Bases de dados de sons

A recuperação de áudio apresenta desafios que são ainda maiores do que os apresentados pela recuperação de imagens. O campo pode ser rigorosamente dividido em recuperação de fala e recuperação de música (embora outros tipos de sons possam também estar presentes em alguns casos). Lu (2001) oferece um levantamento conciso e útil desse campo, embora esteja agora um pouco desatualizado, pois os novos progressos ocorrem muito rapidamente.

Em virtude de uma trilha sonora longa provavelmente apresentar vários componentes de áudio — fala, música e, possivelmente, outros sons (por exemplo, gritos de animais ou ondas lambendo a praia) — o primeiro passo consiste em classificar os vários componentes, e Lu descreve métodos que podem ser usados para se conseguir isso automaticamente.

Os primitivos sistemas de reconhecimento de fala somente podiam funcionar com vocabulários limitados e um número limitado de falantes, porém, desde então, deu-se um notável avanço. Os sistemas atuais são preparados mediante a gravação de seqüências de falas de um grande número de falantes. Da fase de

preparação [*training*] resultam vários produtos, dos quais o mais importante é um dicionário de palavras com suas pronúncias possíveis. Uma nova amostra da fala gravada é comparada com este dicionário e a seqüência de palavras que apresentará a melhor coincidência será emitida como texto gravado. Esta explicação está um tanto simplificada (em primeiro lugar, a unidade de fala usada para comparação está em nível inferior ao da palavra — um fonema) mas serve como idéia geral. Aplicam-se os sistemas de reconhecimento de fala para converter a palavra falada em texto que pode ser processado do mesmo modo que outro texto o é para fins de recuperação. Quer dizer, é possível extrair palavras/expressões que funcionarão como termos de indexação ou fazer buscas no texto inteiro com o emprego dos tipos de procedimento descritos no capítulo 14.

O desempenho dos sistemas de reconhecimento de fala varia segundo alguns fatores, tais como a matéria falada (variando, por exemplo, de números a notícias gerais), seja a fala que resulta da leitura ou de uma conversa espontânea, e o tamanho do vocabulário envolvido. Lu (2001) salienta que o reconhecimento de algarismos pode ser superior a 99%, mas que o reconhecimento de uma conversa telefônica comum pode cair para 50%.

Os problemas da recuperação de documentos falados foram enunciados, de modo muito sucinto, por Wechsler et al. (2000), da seguinte forma:

> O principal problema quando se aplica o reconhecimento de fala à recuperação de documentos falados está na exatidão do resultado do reconhecimento. O reconhecimento automático de fala é uma tarefa difícil e, por conseguinte, seus resultados muitas vezes contêm grande quantidade de *erros de reconhecimento*. A precisão do reconhecimento depende principalmente da: 1) quantidade e qualidade dos *dados acústicos de preparação* [*training data*], 2) quantidade e gênero dos diferentes falantes, 3) quantidade de unidades a serem reconhecidas, e 4) do ambiente de gravação dos documentos falados. Ademais, não há pausas acústicas entre palavras na fala contínua, ao contrário dos espaços em branco num texto.
>
> Os erros de reconhecimento normalmente degradam a eficácia de um sistema de recuperação de documentos falados. São estratégias para superar tal problema: 1) melhorar a precisão do reconhecimento de fala, o que requer enorme quantidade de dados de preparação e tempo, e/ou 2) desenvolver métodos de recuperação que sejam mais tolerantes a erros (p. 173-174).

Um método consiste em desenvolver um reconhecedor de fala que possua um grande vocabulário. Este é empregado para converter a fala em texto que possa então ser manipulado com métodos de recuperação convencionais. Isso exige um investimento muito alto na preparação do dispositivo para reconhecer palavras faladas por diferentes indivíduos, o que implica sua limitação a um domínio ou aplicação restrita (por exemplo, prontuários médicos de pacientes).

Uma abordagem alternativa é passar para um nível inferior ao da palavra e reconhecer e transcrever sons (fonemas). O reconhecimento de fonemas exige menos preparo e, como são unidades mais básicas do que as palavras, torna-se possível ter um vocabulário ilimitado. Os documentos falados serão indexados

e pesquisados sob fonemas, o que equivale aproximadamente ao dispositivo de recuperação de textos que segmenta palavras em bigramas ou trigramas com a finalidade de buscar eficiência. No entanto, os fonemas não são realmente partes de palavras porque, na fala, as palavras freqüentemente fluem juntas, de modo que as unidades reconhecidas são seqüências de fonemas. Isto é, o documento falado é transformado em seqüências fonêmicas, bem como a consulta empregada para interrogar a base de dados. O trabalho de Wechsler et al. (2000) é característico das pesquisas atuais sobre recuperação de seqüências fonêmicas.

É natural que os programas de processamento de fala tenham melhor desempenho no reconhecimento de palavras existentes ('*in-vocabulary*') num *corpus* de preparação [*training corpus*] do que no reconhecimento de palavras não encontradas antes ('*out-of-vocabulary*'). Srinavasan e Petkovic (2000) explicam:

> Um conhecido problema na recuperação de documentos falados é o conceito de termos presentes no vocabulário (*in-vocabulary terms*) e termos ausentes do vocabulário (*out-of-vocabulary terms*). Vocabulário é um conjunto de palavras que um mecanismo de reconhecimento de fala emprega para traduzir fala em texto. Como parte do processo de decodificação, esse mecanismo compara os sons da fala de entrada com as palavras existentes no vocabulário. Portanto, somente as palavras presentes no vocabulário serão reconhecidas. É freqüente uma palavra ausente do vocabulário ser reconhecida, erradamente, como uma palavra ali presente que é foneticamente similar a uma palavra ausente do vocabulário (p. 81).

A decomposição de vocábulos em subvocábulos, isto é, fonemas, normalmente melhora o reconhecimento de palavras presentes no vocabulário, embora não necessariamente o de palavras ausentes do vocabulário. Os efeitos de palavras ausentes do vocabulário na recuperação de documentos falados foram estudados por Woodland et al. (2000).

Brown et al. (2001) relata índices de erros de palavras de 28% no caso de conversas telefônicas de um único falante, e um índice de cerca de 19% na fala preparada (ou seja, não espontânea) de um locutor de notícias em estúdio. Os autores informam que os erros de palavras variaram de 35 a 65% no caso de 'dados de fala do mundo real', a depender de certos fatores, como ruído de fundo, acústica deficiente e participação ou não de falantes nativos. Embora a redundância compense alguns erros, é provável que a recuperação fique bastante prejudicada no caso de 'áudio do mundo real'. Brown et al. relatam valores de revocação da ordem de 26% e valores de precisão por volta de 17% para esse tipo de aplicação, embora resultados muito melhores (por exemplo, precisão de 60 a 70%) sejam alcançados em coleções de teste menores com índices de erros de palavras na faixa de 10 a 30%.

Apesar dos notáveis progressos alcançados no reconhecimento de fala, ainda ocorrem erros de transcrição em proporção séria. Como salientam Moreno et al. (2002):

> Os sistemas de recuperação devem compensar os 20 a 30% de índice de erros de

> palavras que normalmente ocorrem quando reconhecedores de fala que trabalham com grandes vocabulários transcrevem áudio sem restrições como noticiários radiofônicos ou fala informal (p. 58-59).

Allan (2002), no entanto, alega que mesmo altos índices de erros de transcrição podem ser aceitáveis em aplicações de recuperação:

> Mesmo com um índice de erros de reconhecimento de 40%, a eficácia de um sistema comum de recuperação de documentos cai apenas 10% (p. 60).

Ele explica que isso se deve a várias razões: 1) palavras não reconhecidas talvez não sejam necessariamente palavras importantes para a recuperação; 2) redundância (se uma palavra não for reconhecida num lugar, poderá ser reconhecida em outro); 3) sinônimos ou parassinônimos da palavra não reconhecida podem ocorrer e ser reconhecidos.

Moreno et al. (2002) oferecem um bom apanhado sobre os atuais recursos para o reconhecimento de fala:

> Os sistemas de reconhecimento de fala baseados em palavras adotam vocabulários preestabelecidos que incluem de 60 000 a 100 000 vocábulos. O sistema não pode, por definição, presumir palavras fora desse vocabulário. Embora um vocabulário de 100 000 palavras inclua a maior parte das palavras faladas, todo documento inclui pequena porcentagem de palavras ausentes do vocabulário que provavelmente são portadoras de conteúdo, e sua não-inclusão prejudicará o desempenho da recuperação.
>
> Para contornar tal problema, o sistema pode adaptar o vocabulário mediante o exame de documentos relativos ao trabalho. Por exemplo, um reconhecedor de fala usado em sessões de tribunais usaria documentos jurídicos para aprender as palavras do dicionário apropriado. Embora esses vocabulários especializados reduzam o número de palavras ausentes do vocabulário, não garantem sua eliminação (p. 59).

E, em seguida, salientam que os sistemas baseados no reconhecimento de subvocábulos oferecem vantagens:

> Ao invés de reconhecer palavras faladas, esses métodos reconhecem unidades subvocabulares — normalmente, fonemas ou sílabas — com as quais todas as palavras são formadas. O sistema de recuperação de informação decompõe os termos de busca em suas seqüências de subvocábulos constituintes, e então examina os termos reconhecidos para localizar seqüências que correspondam à unidade de busca (p. 59).

Singhal e Pereira (1999) fizeram experiências com a 'expansão de documentos' para compensar erros de transcrição na recuperação da fala. O método deles inclui a expansão de um texto transcrito mediante o acréscimo de palavras de alta freqüência que ocorrem em textos 'relacionados', compensando, graças à redundância, palavras perdidas na transcrição. Parece ser um método muito trabalhoso.

Brown et al. (2001), cujo trabalho também traz uma útil visão da tecnologia de reconhecimento de fala, descreve pesquisas da IBM sobre aplicações de 'mineração de fala'. Uma delas trata de um agente inteligente que captura os debates travados em reunião de negócios ou de pesquisa e "periodicamente torna-se um

participante ativo [...] sempre que encontra informação que identifica como altamente pertinente aos debates em curso". Por exemplo, a ocorrência nos debates do nome de um funcionário pode disparar uma busca nos registros funcionais, a fim de recuperar e tornar disponíveis informações, como endereço, telefone, grupo onde esteja lotado, responsabilidades, experiência. Outras instituições também fizeram pesquisas sobre tecnologia de apoio a reuniões. Brown et al. também descrevem pesquisas sobre a mineração de chamadas de televendas.

As pesquisas sobre recuperação de documentos falados são hoje facilitadas pela existência de uma base de dados de documentos falados no ambiente TREC (Text Retrieval Conferences) (ver capítulo seguinte). O *corpus* TREC 7 consistia em cerca de 100 horas de noticiários radiofônicos, somando cerca de 3 000 notícias. Os grupos de pesquisas participantes trabalharam com transcrições desse *corpus*, de diferentes qualidades, inclusive uma preparada por seres humanos e considerada perfeita, uma preparada por um sistema de reconhecimento de fala com cerca de 35% de índice de erros de palavras, e outra com um índice de erros por volta de 50%. Os grupos participantes testaram seus métodos de recuperação em 23 tópicos pré-selecionados de cada transcrição (Voorhees e Harman, 1999).

As pesquisas sobre interfaces de fala em aplicações de recuperação remontam a vários anos (ver, por exemplo, Smith et al., 1989). Abordagens mais modernas são exemplificadas pelo trabalho de Feder e Hobbs (1995). Ao analisar o emprego da fala humana para alimentação de dados em computador, Shneiderman (2000) apresenta motivos pelos quais as limitações do ser humano (por exemplo, fadiga, impaciência, dificuldades de corrigir erros) seriam mais importantes do que as limitações tecnológicas.

Métodos modernos para sintetizar e arquivar sons eletronicamente tornam disponível grande quantidade de sons (por exemplo, para músicos), mas a recuperação de um som específico desse arquivo constitui grande problema. Feiten e Günzel (1994) descrevem uma abordagem da indexação e recuperação de sons por meio de redes neurais. O índice de recuperação é criado automaticamente. A capacidade de reconhecer e rotular (isto é, indexar) sons automaticamente tem muito em comum com o processamento necessário para reconhecer imagens automaticamente. Como salientam Picard e Minka (1995), tanto há uma 'textura de sons' quanto uma textura de imagens. Assim, seria possível desenvolver técnicas para identificar automaticamente certos sons (um sino a badalar, água a correr, aplausos) mediante alguma forma de cotejo de padrões (sonoros). A recuperação de sons é analisada em trabalho de Blum et al. (1997), que descrevem um 'navegador de sons' desenvolvido para possibilitar buscas difusas em bases de dados de áudio. Os recursos incluem consulta por exemplo (isto é, 'encontre sons semelhantes a ...').

Recuperação de música

O objetivo das abordagens modernas da recuperação de música é "responder

consultas de música formuladas musicalmente" (Downie e Nelson, 2000) — isto é, permitir que seja feita uma busca baseada numa entrada musical (por exemplo, cantada ou cantarolada).

A história da recuperação de informação musical remonta à década de 1960, mas a maioria dos progressos alcançados se deu a partir da década de 1990. Encontra-se condensada nos anais de três simpósios internacionais sobre a matéria, realizados em 2000, 2001 e 2002. Os trabalhos de 2000 estão disponíveis no sítio <http://ciir.cs.umass.edu/music2000/papers.html> e os de 2001 em <http://ismir2001.indiana.edu/papers.html>. Um objetivo importante desses simpósios é o desenvolvimento de uma coleção-padrão de música, consultas e avaliações que possam ser usadas para comparar diferentes métodos, de modo muito parecido com a forma como funcionam as conferências TREC.

A recuperação de música é mais complexa do que a de fala. Lu (2001) divide o campo em: 1) música estruturada ou sintética, e 2) música baseada em amostras [*sample-based*]. Na primeira, as notas musicais são gravadas como algoritmos e linguagens de controle, que torna o cotejo com as consultas (na forma de uma seqüência de notas) relativamente fácil, pelo menos no caso de coincidência exata. A detecção de passagens de música 'semelhante' é mais complicada.

Muito mais complexa é a recuperação de música que não esteja gravada em formato estruturado. Lu (2001) refere-se a essa música como 'baseada em amostras' porque ela implica o reconhecimento e extração de *samples* [amostras] musicais. Ele identifica duas abordagens de indexação/recuperação. A primeira baseia-se na extração de 'características acústicas' (como audibilidade, tom, brilho, largura de banda e harmonicidade) e que podem ser calculadas para cada 'quadro' da composição gravada. Uma composição musical, usada como consulta (normalmente uma forma cantarolada), é reduzida às mesmas características, o que permite busca com base numa comparação de padrões. Na segunda abordagem, a indexação e a recuperação baseiam-se no tom. Para cada nota extrai-se ou se calcula o tom. Cada tom pode ser representado como uma mudança (para cima, para baixo ou similar) relativa ao precedente, e assim a composição musical (ou composição de consulta) é representada por meio de uma seqüência de símbolos que representam essas alterações de tom. Alternativamente, cada nota musical pode ser representada por um valor de tom selecionado de um conjunto de valores-'padrão' de tom numa base de maior coincidência. De novo, a composição musical será representada por uma seqüência de caracteres que representam o valor do tom.

Lippincott (2002) nos oferece uma descrição bem útil e concisa daquilo que as atuais abordagens da recuperação de música estão procurando realizar:

> Antigamente, os usuários que procuravam informações sobre música voltavam-se para fontes impressas que continham metadados registrados à mão e ordenados por título, compositor e outras categorias. Obviamente, os métodos de acesso refletiam técnicas de recuperação da época, baseadas em material impresso para recuperação de informação bibliográfica, e também pressupunham algum conhecimento musical prévio ou a presença de um bibliotecário. Grande parte das pesquisas atuais sobre recuperação automatizada de informação musical baseia-se em caracterizações da própria música, ao invés de informações sobre ela. Por exemplo, ao invés de solicitar uma busca por título da composição, o usuário entra com uma consulta no formato de áudio e recupera resultados similares a essa consulta. As implicações para os usuários comuns de sistemas de recuperação de música baseados em conteúdo são importantes, pois não é preciso o conhecimento bibliográfico prévio de uma composição musical; ao contrário, bastará, para fins de recuperação, um trechinho de música a fluir na mente do usuário (p. 137).

Este trabalho é uma ótima síntese de vários métodos que vêm sendo pesquisados.

Liu e Tsai (2001) salientam que:

> A maneira mais direta de que um usuário leigo dispõe para consultar as bases de dados de música é cantarolar uma composição como uma consulta-exemplo para recuperar objetos musicais similares (p. 506).

Um dos problemas, porém, é a grande diferença de extensão entre esse tipo de consulta-exemplo e uma composição musical: uma consulta feita com uma música cantarolada normalmente dura alguns segundos, enquanto uma música popular comum dura cerca de cinco minutos. Os autores descrevem um método experimental em que o cotejo se torna mais eficiente mediante o seqüenciamento de uma composição musical em 'fases' que têm aproximadamente a mesma extensão de uma consulta feita com música cantarolada.

Na indexação e recuperação de música, é preciso distinguir entre música monofônica (nenhuma nota começa até que a nota atual tenha terminado de soar) e a música polifônica (uma nota pode começar antes que a anterior termine). A música polifônica é mais comum, porém mais complexa para as operações de indexação e recuperação. Pickens (2001) descreve os problemas de seleção de características para indexação e recuperação de música polifônica.

Diversas abordagens da indexação e recuperação de música polifônica foram apresentadas. Ver, por exemplo, Dovey (2001) e Doraisamy e Rüger (2001).

Downie e Nelson (2000) descrevem um método de recuperação de música baseado no tom, especificamente a diferença entre dois tons, conhecida como 'intervalo'. As melodias de uma coleção de canções folclóricas foram "convertidas em representações de um único intervalo de melodias monofônicas". Estas foram então fragmentadas em subseções designadas 'n-gramas', que são usadas para formar 'palavras musicais'. Isso permite uma abordagem da recuperação que se assemelha à busca de palavras na recuperação de textos e possibilita que seja aplicado um sistema de processamento baseado em textos (o SMART de Salton) que permite recuperação em ordem de provável relevância.

É possível também usar entrada em formato de áudio para buscas em base de dados de partituras musicais. McNab et al. (2000) descrevem um método para recuperar partituras de uma base de dados em resposta a 'poucas notas entoadas

ou cantaroladas num microfone'. A interface adotada transcreve a entrada acústica em notação musical comum que pode ser usada para cotejo seqüencial e recuperação de música em ordem de provável relevância. Seu protótipo 'prova de conceito' foi testado numa base de dados de canções folclóricas. Concluíram que:

> Não é uma empresa simples fazer buscas em grandes bases de dados de música e recuperar itens em que ocorra um determinado tema ou seqüência de notas, tendo em vista principalmente as imprecisões que ocorrem quando as pessoas entoam melodias, mas isso está com certeza ao alcance da tecnologia atual (p. 113).

Byrd e Crawford (2002) fizeram uma revisão do estado atual dos conhecimentos a respeito da indexação e recuperação de música e concluíram que o progresso alcançado nessa área foi muito limitado:

> Apesar de expressivo número de projetos de pesquisa haver se voltado para a recuperação de informação musical, nas últimas três décadas, esse campo ainda está muito imaturo. Poucos dizem respeito à música complexa (polifônica); os métodos de avaliação ainda estão numa etapa de desenvolvimento muito primitiva; nenhum dos projetos enfrenta o problema de bases de dados que são, realisticamente, de grande escala. Muitos dos problemas a serem enfrentados se devem à natureza da própria música. Entre eles estão as questões ligadas à percepção humana e à cognição da música, especialmente no que tange à reconhecibilidade da frase musical [...] e o pressuposto comum de que buscas sobre o tom (ou contorno do tom) provavelmente bastariam para atender a todas as finalidades [...] talvez seja verdadeiro para a maior parte da música monofônica (de uma só voz), mas é certamente inadequado para música polifônica (de muitas vozes). Mesmo no caso monofônico pode levar a resultados equivocados. O fato, há muito admitido em projetos que dizem respeito à música monofônica, de que uma passagem reconhecível normalmente não é idêntica ao padrão de busca significa que quase sempre é necessária uma coincidência aproximada, mas também isso se torna seriamente complicado pelas demandas da música polifônica. Quase todos os métodos de recuperação da informação de textos apóiam-se na identificação de unidades aproximadas de sentido, isto é, palavras. Um problema fundamental da recuperação da informação em música está em que é extremamente difícil, talvez impossível, localizar essas unidades (p. 249).

Sistemas multimídias

Até agora este capítulo tratou da recuperação de imagens e da recuperação de sons. No entanto, também estão em curso pesquisas sobre problemas de indexação e recuperação relativos a apresentações verdadeiramente multimídias, como as transmissões de televisão.

Um sistema de indexação de multimídias descrito por Kubala et al. (2000) processa a linguagem falada produzida por fontes de áudio e vídeo, como os noticiários de televisão. O protótipo desse sistema possui recursos para sumarização* e indexação. O autor descreve o primeiro deles da seguinte forma:

* Os problemas implicados na criação automática de resumos de diálogo falado são bem analisados por Zechner (2001).

> A sumarização é uma representação estrutural do conteúdo em linguagem falada que é muito poderosa e flexível como índice para gerenciamento de informações baseadas em conteúdo. Este resumo, que é produzido automaticamente pelo sistema, inclui características extraídas, como nomes de pessoas, lugares e organizações mencionados no transcrito, bem como as identidades e localizações dos falantes na gravação (p. 49).

O fluxo contínuo de palavras é automaticamente segmentado em 'passagens que são tematicamente coerentes' e cada passagem é indexada mediante a atribuição automática de 'rótulos tópicos' extraídos de um conjunto preestabelecido de mais de 5 000 desses rótulos. Estes são classificados em ordem de probabilidade de adequação e são atribuídos a cada passagem os rótulos de classificação mais alta.

Importante projeto de indexação e recuperação de multimídia é a Informedia Digital Video Library da Carnegie Mellon University. Wactlar et al. (2000) assim descrevem seus recursos:

> [...] emprega exclusivamente fala, imagem e compreensão da linguagem natural integradas para processar transmissões de vídeo. [...] A fim de possibilitar este acesso ao vídeo, são geradas, por meio do sistema de reconhecimento de fala Sphinx, da Carnegie Mellon University, transcrições rápidas, de alta precisão e automáticas, de noticiários de televisão, sendo incorporadas legendas fechadas onde estiverem disponíveis. O processamento da imagem determina limites de cenas, reconhece rostos e permite comparações de semelhança de imagens. O texto visível na tela é reconhecido por meio de reconhecimento de caracteres ópticos de vídeo e pode ser pesquisado. Tudo é indexado numa biblioteca digital de vídeo pesquisável, onde os usuários podem formular consultas e recuperar, como resultado, notícias relevantes [...]
>
> O sistema Informedia permite recuperação da informação tanto no domínio da linguagem falada quanto no domínio do vídeo ou imagem. As consultas em busca de notícias relevantes podem ser feitas por meio de palavras, imagens ou mapas. Rostos são detectados no vídeo e podem ser pesquisados. Resumos informativos podem ser exibidos com informações variáveis, tanto visual quanto textualmente. Os resumos de textos são exibidos para cada notícia por meio de tópicos e títulos. São oferecidos resumos visuais por meio de imagens miniaturizadas [*thumbnails*], tiras de filme [*filmstrips*] e sínteses [*skims*] dinâmicas de vídeo (p. 42-43).

Wactlar et al. afirmam ser possível um índice de erros inferior a 20% no reconhecimento de fala e que a transcrição de um noticiário pode aparecer na base de dados duas horas e meia depois de haver sido transmitido.

Brown et al. (2001) oferecem mais esclarecimentos:

> O projeto de pesquisa Informedia criou uma biblioteca digital de um milhão de megabytes em que descritores obtidos automaticamente para vídeo são utilizados na indexação, segmentação e acesso ao conteúdo da biblioteca. Combina reconhecimento de fala, processamento de imagens e técnicas de compreensão da linguagem natural para o processamento automático de vídeo, a fim de produzir uma síntese [*skim*] visual, que diminui o tempo de visualização sem perda de conteúdo. Oferece três maneiras de visualização dos resultados das buscas: quadros-pôster [*poster frames*], tiras de filme e sínteses. A visualização em quadros-pôster apresenta os

> resultados da busca em formato de quadros-pôster, em que cada quadro representa um 'parágrafo' de vídeo. A visualização em tiras de filme reduz a necessidade de visualizar cada parágrafo de vídeo em sua totalidade ao oferecer páginas de *storyboard* para rápida visualização. As subseções mais relevantes do parágrafo de vídeo são exibidas como cenas-chave e as palavras-chave são nitidamente marcadas. A recuperação combinada de palavras e fones também foi investigada no projeto Informedia, onde se utilizou um índice invertido para transcrição fonética, que inclui subseqüências fonéticas de três a seis fones. Na recuperação o índice de documentos com palavras e a transcrição fonética são pesquisados em paralelo e os resultados são fundidos. Experiências com um *corpus* de cerca de 500 notícias dos noticiários da ABC e da CNN (Cable News Network), com o emprego de índices combinados de palavras e fones, resultou numa precisão média de 0,67 com um desempenho global de 84,6% do de um sistema de recuperação de texto completo. No caso, porém, de áudio do mundo real com alto índice de erros de palavras de 70–80%, registrou-se uma queda drástica da precisão e revocação para 0,17 e 0,26, respectivamente (p. 989-990).

Os recursos de buscas de imagens do sistema Informedia incluem detecção de cor (o usuário especifica cores e regiões de interesse a serem procuradas entre as imagens). Ver Wactlar et al. (1999).

Patel e Sethi (1996) descrevem métodos que desenvolveram para classificar segmentos de filmes cinematográficos mediante processamento de áudio. De início, o sistema somente podia identificar categorias genéricas (como, por exemplo, 'musical'), porém os autores sugerem que ele poderia ser mais aprimorado de modo a identificar especificamente tipos de cenas (cena de ação, cena de dança, cena romântica, e assim por diante). Posteriormente (Patel e Sethi, 1997) estenderam sua pesquisa à identificação dos falantes (por exemplo, atores em trechos em vídeo de filmes).

Adami et al. (2001) propõem um sistema que oferece acesso a documentos multimídias por meio de ferramentas análogas às de um livro impresso: uma descrição hierárquica do conteúdo do item (similar a uma página convencional de sumário) adequada para pesquisa, e um 'índice analítico' baseado em palavras-chave (análogo ao índice do final de um livro). Sua pesquisa tem por objetivo produzir essas ferramentas de modo automático, e mostram um exemplo baseado na análise de um jogo de futebol.

Gauvain et al. (2001) descrevem um sistema de partilhamento e transcrição automáticos de transmissões de televisão e rádio. Segmentos de não-fala das transmissões são identificados e removidos (automaticamente) e os segmentos restantes são agrupados e rotulados de acordo com a largura de banda e o gênero. Um 'reconhecedor de fala contínua, independente de falante e de vocabulário extenso' é empregado para preparar as transcrições. Afirma-se a ocorrência de uma média de erros de palavras de 20%.

Conclusões

Houve muito progresso na indexação e recuperação de imagens na última década, e algum avanço se deu na indexação e recuperação de sons. Os inúmeros estudos realizados sobre a reação do observador às imagens são, contudo, de qualidade variável. Alguns são úteis. Mas outros, especialmente os realizados como pesquisa para redação de teses, deixam muito a desejar. Embora sejam apresentadas minuciosas análises de dados, um número muito grande desses estudos chega a conclusões que não são nem mesmo abordadas pelos dados coletados, dando a impressão de que as conclusões foram definidas antes da realização de qualquer estudo.

Os campos da recuperação de imagens e sons atraíram muitos pesquisadores que não possuíam qualquer experiência anterior com a recuperação de textos. Disso resultou o surgimento de uma nova terminologia para idéias muito antigas, o que, na realidade, é muito lamentável. Um exemplo primoroso é o uso da palavra 'anotação' para designar a atribuição de um rótulo verbal a uma imagem — ou seja, sua indexação (ver, por exemplo, Picard e Minka, 1995).

Naturalmente, os futuros desenvolvimentos na recuperação do discurso falado dependem em muito dos progressos que ocorrerem no campo geral da tecnologia da fala. As revistas mais populares do ramo tendem a ser exageradamente otimistas quanto às futuras possibilidades. Por exemplo, a afirmativa de Flynn (1993):

> No final da década, os sistemas de reconhecimento de fala permitirão a você falar naturalmente, com um vocabulário virtualmente ilimitado (p. 29)

era totalmente irreal da maneira como foi formulada.

Haas (1996), citando Rudnicky, levanta uma questão importante que é pertinente às perspectivas nessa área:

> Há uma diferença entre reconhecimento de fala e compreensão de fala: o reconhecimento de fala requer que um sistema identifique as palavras numa expressão oral, enquanto a compreensão de fala requer que um sistema também trate dos problemas ligados à compreensão da linguagem natural, como anáfora, elipse e outros fenômenos do discurso. O reconhecimento de fala é útil para tarefas estruturadas, como entrada de dados e emissão de comandos simples, mas um diálogo, de qualquer tipo, exige compreensão de fala (p. 98).

A compreensão da fala humana pelo computador não é uma perspectiva que esteja presente no horizonte imediato.

Mesmo no seio da comunidade de pesquisadores desse campo, há uma ampla divergência de opiniões quanto ao que foi alcançado pela tecnologia de reconhecimento de fala e o que poderá suceder no curto prazo. Levinson (1995), por exemplo, acredita que ainda se passará muito tempo antes que surjam sistemas de real valor comercial:

> A opinião da maioria assegura que logo os melhoramentos técnicos tornarão o reconhecimento de fala baseado em grandes vocabulários comercialmente viável para aplicações específicas. Minha previsão [...] é que os melhoramentos técnicos surgirão de modo penosamente lento, mas que dentro de 40 a 50 anos o reconhe-

> cimento de fala com níveis de desempenho dos seres humanos estará onipresente. Isto é, progressos técnicos incrementais resultarão, em curto prazo, numa tecnologia frágil de valor comercial relativamente modesto em mercados muitos especiais, enquanto importantes avanços tecnológicos resultantes de uma verdadeira mudança de paradigma na ciência subjacente possibilitarão às máquinas mostrar níveis humanos de competência na comunicação por meio da linguagem falada. Isso, por sua vez, resultará num vasto mercado de incalculável valor comercial (p. 9954).

No entanto, Srinavasan e Brown (2002) frisam que, embora a tecnologia da fala estivesse lenta para encontrar aplicações comerciais, parece que agora está pronta para decolar comercialmente:

> A conectividade da Rede, a tecnologia sem fio e os dispositivos portáteis de mão — combinados com o reconhecimento eficaz de fala baseado na gramática [...] — podem finalmente levar o reconhecimento de fala a ter a importância de um mercado de massa (p.38).

Afirmativas exageradas também ocorrem no campo da recuperação de imagens. Muitos pesquisadores nesse terreno são completamente ingênuos em suas crenças e expectativas. Para citar somente um exemplo, Gupta e Jain (1997), num estudo panorâmico da recuperação de imagens, útil por outros motivos, estimula-nos da seguinte forma:

> Os usuários podem agora extrair, armazenar e recuperar conteúdo informacional 'baseado em imagens' — metadados e atributos visuais — de mídia visual de modo tão fácil quanto a procura de documentos textuais (p. 71).

Aqueles que vimos trabalhando nessa área há mais de 40 anos sabemos que a recuperação de documentos textuais está muito longe de ser fácil em bases de dados de porte significativo.

É importante admitir que as pesquisas sobre recuperação de imagens ou sons dependem muito mais das técnicas de indexação automática do que da indexação feita por seres humanos. Por isso, as abordagens que serão objeto dos dois próximos capítulos relacionam-se bem de perto com o conteúdo deste.

CAPÍTULO 14

Buscas em textos

A aplicação de computadores à recuperação de informações, que teve início na década de 1950, possibilitou a realização de buscas em textos em formato eletrônico, sem que houvesse a necessidade de aplicar qualquer modalidade de indexação a este texto: o programa utilizado na recuperação procura determinadas palavras, ou combinações de palavras, no próprio texto, onde as palavras escolhidas por quem faz a busca são indicativas daquilo que o texto está examinando. As buscas feitas em textos pelo computador podem ser denominadas 'buscas em textos' ou 'buscas em linguagem natural'. O texto onde são feitas as buscas pode ser o conteúdo completo de uma publicação (artigo, relatório, ou até um livro) ou parte dela: o resumo, extrato ou apenas o título. As buscas feitas num texto integral são às vezes denominadas 'buscas em texto completo'.

A viabilidade de buscas em textos cresceu notavelmente ao longo dos anos, na medida em que aumentou o potencial dos computadores, os custos de processamento e armazenamento diminuíram, e um volume cada vez maior de textos tornou-se disponível em formato eletrônico, em grande parte como subproduto de várias formas de publicação. O desenvolvimento da Rede, que torna acessível enorme quantidade de textos a um imenso número de usuários, tornou rotineira, ao invés de excepcional, a busca em textos. Por causa disso, o interesse por métodos de buscas em textos aumentou notavelmente na última década, tanto na comunidade de pesquisa quanto em setores governamentais e comerciais.

Essa área de buscas em textos vem avançando desde 1991, graças ao programa TIPSTER e a várias outras atividades com ele relacionadas. O TIPSTER foi uma iniciativa da Defense Advanced Research Projects Agency (DARPA), em colaboração com o National Institute of Standards and Technology, outros órgãos governamentais e várias empresas comerciais. O programa teve vários componentes, dos quais o mais pertinente ao conteúdo deste capítulo foram as conferências anuais intituladas Text Retrieval Conferences (TRECs), de que foram realizadas 11 até 2002. As atividades das TRECs impulsionam o estado de desenvolvimento da área, ao permitir que diferentes grupos de pesquisas testem e comparem seus programas de recuperação em condições controladas (bases de dados, consultas e avaliações de relevância mantidas constantes). Outros componentes do TIPSTER serão focalizados no capítulo seguinte.

Na realidade, este capítulo e o próximo se inter-relacionam tão de perto que deverão, de fato, ser lidos como uma unidade. Às vezes, foi uma decisão um

tanto arbitrária decidir quanto ao que incluir neste capítulo e ao que passar para o seguinte.

Os procedimentos modernos de processamento de textos alegam que aplicam técnicas oriundas de pesquisas em inteligência artificial, e a expressão 'processamento inteligente de textos' é às vezes empregada para designar esse tipo de processo (ver, por exemplo, Jacobs, 1992c).

Este capítulo passará em revista os méritos relativos das abordagens da recuperação da informação baseadas em textos (linguagem natural) e em vocabulários controlados, fará um levantamento do desenvolvimento das buscas em textos desde a década de 1950 e terminará com considerações sobre os atuais recursos nesta área.

Considera-se a expressão *linguagem natural* como sinônimo de 'discurso comum', isto é, a linguagem utilizada habitualmente na escrita e na fala, e que é o contrário de 'vocabulário controlado'. No contexto da recuperação da informação, a expressão normalmente se refere às palavras que ocorrem em textos impressos e, por isso, considera-se como seu sinônimo a expressão 'texto livre'. Um texto livre pode consistir em:

1. o título,
2. um resumo,
3. um extrato, ou
4. o texto integral de uma publicação.

Embora 'texto livre' se refira usualmente a uma parte integral de um texto, esta expressão é também empregada para designar palavras ou expressões extraídas do texto por um indexador humano (ou por programa de computador) e incluídas num registro bibliográfico que representa o texto. Em alguns casos, os termos assim extraídos são acrescidos aos títulos de itens indexados, formando títulos 'expandidos' ou 'enriquecidos'.

Um pouco de história

Os métodos 'modernos' que visam ao uso da linguagem natural na recuperação da informação remontam ao sistema Uniterm descrito por Taube em 1951. Os princípios do sistema Uniterm despertaram atração imediata: o conteúdo temático dos documentos podia ser representado adequadamente por meio de palavras simples (unitermos) extraídas do texto dos documentos por indexadores com um nível de especialização relativamente baixo. Escritos à mão ou datilografados, os números dos documentos eram 'lançados' em fichas projetadas para esse fim, cada uma representando um único termo, e as buscas eram feitas comparando-se os números em duas ou mais fichas (de modo muito parecido com um moderno sistema em linha que compara listas de números associados a termos).

Taube teve considerável influência sobre o desenvolvimento de sistemas de recuperação da informação na década de 1950. Infelizmente, todavia, o sistema Uniterm veio a ser na prática menos atraente do que parecera à primeira vista.

Padecia de todos os problemas para cuja solução os vocabulários controlados foram criados. Conteúdos temáticos que apresentavam relações muito próximas entre si apareciam sob diferentes unitermos, e uma busca exaustiva sobre um assunto exigia que se imaginassem todas as formas como esse assunto estaria representado no texto, o que nem sempre era uma tarefa fácil. Esses problemas acarretaram um retorno aos vocabulários controlados e ao desenvolvimento do tesauro para a recuperação da informação (Holm & Rasmussen, 1961).

Além dos problemas de ordem terminológica, o sistema Uniterm também padecia de limitações mecânicas. Quem fosse fazer uma busca somente poderia cotejar com facilidade duas fichas de cada vez. Assim, uma busca sobre *A* em relação a *B*, onde *A* estivesse representado por quatro unitermos e *B* por dez, exigiria que fossem feitas 4×10 cotejos de fichas separadamente. Embora isso fosse possível, tratava-se de uma tarefa enfadonha e demorada. Além disso, ainda que a relação booleana *e* (que envolve a comparação de números) seja fácil de fazer mediante a manipulação de fichas do sistema Uniterm (ou fichas *peek-a-boo*), fica muito difícil em sistemas manuais desse tipo realizar uma busca booleana usando *ou* (que envolve a fusão de listas) e principalmente combinar (*e*) conjuntos de termos numa relação *ou*. Tais manipulações de termos são, naturalmente, comuns em sistemas informatizados. O computador, portanto, soluciona os problemas 'mecânicos' acarretados pela manipulação de inúmeros termos não-controlados, mas não resolve, por si mesmo, os problemas intelectuais criados pela inexistência de controle do vocabulário.

Todavia, quando os computadores foram inicialmente aplicados à recuperação da informação, em escala importante, em fins da década de 1950 e início da década de 1960, reconheceu-se que as buscas em textos, e mesmo buscas em textos integrais, haviam se tornado uma possibilidade sedutora. Ao se estudar a história dos sistemas informatizados de recuperação da informação, reconhecem-se duas linhas principais de desenvolvimento. Uma delas tem sua origem nos grandes sistemas, desenvolvidos por certas instituições como a National Library of Medicine (NLM), o Department of Defense (DOD) e a National Aeronautics and Space Administration (NASA), que funcionavam com base em termos de indexação extraídos de um vocabulário controlado e atribuídos aos documentos por indexadores humanos. A outra linha de desenvolvimento teve seu início no campo do direito, e envolvia a colocação de textos completos (por exemplo, leis) em formato eletrônico e a utilização do computador para fazer buscas de palavras ou combinações de palavras nesses textos. Trabalhos dessa natureza antecederam, na realidade, o desenvolvimento de tesauros e o surgimento dos grandes sistemas baseados na indexação feita por seres humanos. A recuperação de textos jurídicos integrais remonta ao trabalho de Horty e seus colaboradores no Health Law Center da University of Pittsburgh (Horty, 1960, 1962, Kehl et al., 1961). Foi no campo jurídico que as técnicas modernas de buscas em texto livre tiveram seu desenvolvimento inicial, e o trabalho pioneiro em Pittsburgh lançou os alicerces dos sistemas posteriores de recuperação de informação jurí-

dica exemplificados por LEXIS e WESTLAW. Myers (1973) apresentou uma útil revisão sobre o estado dos conhecimentos relativos à busca em textos jurídicos por computador. Embora antigo, continua sendo um bom relato acerca dos princípios básicos. Dabney (1986) serve como uma atualização.

A distinção entre os sistemas baseados essencialmente em vocabulários controlados e registros de indexação criados por seres humanos (muitas vezes equivocadamente denominados sistemas 'bibliográficos') e os sistemas baseados em buscas no texto tem se tornado cada vez mais difusa com o passar dos anos. Gradualmente, os sistemas 'bibliográficos' foram permitindo a busca de palavras que ocorriam nos títulos e, depois, nos resumos, enquanto alguns dos sistemas de texto integral acrescentavam termos de indexação atribuídos por seres humanos a fim de melhorar o acesso, e algumas bases de dados (por exemplo, INSPEC) foram projetadas, desde o início, para incluir tanto termos controlados quanto 'palavras-chave' não controladas. Na medida em que um número cada vez maior de textos tornou-se disponível em formato eletrônico, como subproduto de atividades de editoração ou disseminação, a busca em textos de resumos passou a ser um lugar-comum, e a busca em textos completos ultrapassa hoje as fronteiras do direito: jornais, revistas de cunho popular, periódicos científicos, enciclopédias e outras fontes encontram-se agora acessíveis em formato de texto completo. Os sítios da Rede da internet consistem majoritariamente em texto, de modo que uma verdade indubitável é que as buscas em textos superam hoje grandemente as buscas que envolvem vocabulários controlados.

As buscas em textos são realizadas de dois modos. No primeiro método, palavras que não sejam comuns são incluídas em arquivos 'invertidos', que mostram, para cada palavra, qual o documento em que ela aparece (e freqüentemente sua posição exata nesse documento). A busca é realizada nesses índices (os quais, no trabalho originalmente desenvolvido em Pittsburgh, eram denominados 'concordâncias') ao invés de ser realizada no próprio texto. A outra alternativa é efetuar uma busca seqüencial no texto, palavra por palavra, sem utilizar qualquer índice. Esta era a técnica comumente empregada para prestar serviços de Disseminação Seletiva de Informações (DSI) a partir de bases de dados, antes de estarem amplamente difundidos os sistemas em linha. Quer dizer, os perfis de interesses de usuários, que se achavam armazenados, eram comparados com atualizações periódicas da base de dados (palavras nos títulos ou resumos). Este método 'caudaloso' de buscas em textos era mais atraente no caso de aplicações de DSI do que em buscas retrospectivas devido a que o volume de texto a ser examinado em qualquer momento é muito menor na DSI. Mais tarde, contudo, foram desenvolvidos computadores especializados que podiam fazer buscas em textos de modo tão rápido que se tornaram bastante viáveis as buscas 'caudalosas' até mesmo em bases de dados muito grandes. Por exemplo, o Fast Data Finder (Yu et al., 1987) alegava que realizava buscas em textos à velocidade de 12,5 milhões de caracteres por segundo, o que equivale a cerca de 12,5 romances de 500 páginas a cada segundo.

Embora o método 'caudaloso' não seja conceitualmente diferente do método de índice invertido, possui, de fato, algumas características melhoradas. Por exemplo, é muito mais fácil realizar buscas com 'fragmentos' de palavras, principalmente seqüências de caracteres que ocorram no meio ou no fim de um vocábulo.

Os mecanismos de busca que foram desenvolvidos na internet funcionam por intermédio da compilação de 'índices' de textos presentes nos vários sítios e não passam de arquivos invertidos convencionais.

Recursos auxiliares de busca

Mesmo nos primórdios das buscas em textos, vários recursos auxiliares já haviam sido desenvolvidos para ajudar quem realizava as buscas. O mais primitivo deles é a apresentação (ou saída impressa) em ordem alfabética das palavras 'significativas' que ocorrem na base de dados, com uma indicação da freqüência com que cada uma delas ocorre. Também era comum algum tipo de indicador de distância entre as palavras (operador métrico). A capacidade de especificar a proximidade entre duas palavras é particularmente útil em buscas em bases de dados de textos completos onde palavras que ocorrem em parágrafos diferentes podem não estar de modo algum relacionadas diretamente entre si.

Talvez o recurso auxiliar mais poderoso das buscas em linguagem natural seja a capacidade de realizá-las em partes de palavras — quer dizer, fazer seu truncamento ou efetuar buscas com fragmentos de palavras. A utilidade das buscas com fragmentos de palavras foi analisada por Williams (1972). Os programas de computador mais flexíveis permitem que se façam buscas com qualquer fragmento: truncamento à direita (por exemplo, todas as palavras iniciadas com 'condens'), truncamento à esquerda (todas as palavras que terminam com 'micina'), truncamento com 'infixos' (especificam-se o começo e o fim mas não o meio da palavra), ou qualquer combinação possível desses recursos (por exemplo, todas as palavras que incluam a cadeia de caracteres 'magnet', independentemente de onde apareça). Embora sejam potencialmente úteis em todos os campos do conhecimento, as buscas com fragmentos de palavras parecem ter mais utilidade em ciência e tecnologia, onde a linguagem costuma ser mais previsível. Em certo sentido, esse recurso permite que se compense a ausência de um vocabulário controlado mediante a formação de classes úteis de palavras numa estratégia. Assim, as buscas com o radical 'condens' provavelmente possibilitarão a recuperação de um grupo de documentos que terão algo a ver com condensadores e condensação; buscas com o sufixo 'micina' resultarão em documentos que tratam de antibióticos; e buscas com 'tri...cobaltato' (infixo não especificado) recuperarão uma família de compostos químicos afins.

As buscas com fragmentos de palavras oferecem alguns recursos do tesauro convencional, mas o fazem quando da saída, ao invés do controle feito na etapa de entrada. Por exemplo, a possibilidade de buscas com os sufixos 'bióticos ou ilina ou micina ou ciclina ou mixina' quase equivale a uma entrada 'antibióticos'

num tesauro convencional que leva a uma lista de termos específicos relativos a antibióticos. O tesauro convencional é um vocabulário pré-controlado, enquanto a formação de classes de palavras ou fragmentos de palavras numa estratégia de busca é uma espécie de processo de 'pós-controle'.

Linguagem natural versus vocabulário controlado: algumas considerações gerais

Alguns fatores importantes que influem no desempenho dos sistemas de recuperação da informação podem ser exemplificados reportando-nos mais uma vez à figura 3. Nela, à esquerda, se encontram três representações em texto livre de um documento (um título e dois resumos de extensão variável), enquanto à direita aparecem dois conjuntos de termos de indexação (cobertura seletiva e exaustiva do conteúdo temático). Os termos foram extraídos do *UNBIS thesaurus* (Nações Unidas, Dag Hammarskjold Library, 1985). Um fator importante que influi no desempenho dos sistemas de recuperação da informação é o número de pontos de acesso providos. Evidentemente, o resumo expandido provê mais pontos de acesso do que o resumo sucinto, o qual, por sua vez, provê mais do que o título. Do mesmo modo, a indexação exaustiva provê um número de pontos de acesso quase três vezes maior do que o provido pela indexação seletiva.

Uma busca em texto que se restrinja apenas ao título provavelmente permitirá que esse item só seja recuperado numa busca sobre o conteúdo temático dominante do documento. À medida que se acrescenta mais texto, o item torna-se recuperável no curso de buscas sobre outros aspectos. O resumo sucinto permitiria recuperação em buscas sobre: ajuda norte-americana, a OLP, o Estado palestino, Israel, ajuda norte-americana a Israel e conferências de paz, enquanto o resumo expandido acrescenta outros pontos de acesso, tais como esforços pela paz e líderes do Oriente Médio. Evidentemente, isso também ocorre na comparação entre indexação seletiva e indexação exaustiva. A indexação seletiva reflete apenas o título do item e não provê pontos de acesso adicionais ao título, e a indexação exaustiva equivale mais ou menos em amplitude ao resumo expandido.

Ao se considerar a recuperabilidade do item apresentado, é a extensão do registro que tem maior importância, e não o tipo de vocabulário. A indexação seletiva, quanto a isso, equivale ao título, enquanto a indexação exaustiva se situa em algum ponto entre os dois resumos na medida em que abrange o conteúdo temático do item. Como o resumo sucinto provê mais pontos de acesso do que o título ou a indexação seletiva, o item que representa será mais recuperável. Do mesmo modo, a indexação exaustiva torna esse item mais recuperável do que o seria numa busca no resumo sucinto, porém menos recuperável do que o seria numa busca no resumo expandido.

Uma base de dados constituída de milhares de itens indexados exaustivamente, como no exemplo da figura 3, provavelmente possibilitará uma revocação muito mais alta do que uma outra que proveja acesso apenas pelos

títulos. Do mesmo modo, uma base de dados constituída de resumos 'expandidos' provavelmente possibilitará revocação mais alta do que uma outra baseada na indexação seletiva ou mesmo, talvez, na indexação exaustiva. Isso, em si mesmo, nada tem a ver com a comparação entre linguagem natural *versus* vocabulários controlados na recuperação da informação, mas diz respeito apenas à extensão do registro no qual se podem realizar buscas.

Vários pesquisadores, infelizmente, deixaram de atentar para a extensão do registro ao compararem a recuperação baseada em texto livre com a recuperação baseada na indexação. Não é de estranhar que a indexação exaustiva resulte em revocação mais alta do que os títulos e isto não prova que a indexação feita por seres humanos seja superior ao texto livre. Não é de estranhar que um resumo extenso resulte em revocação mais alta do que a indexação seletiva e isto não prova que o texto livre seja superior à indexação feita por seres humanos.*

Os resumos, todavia, freqüentemente proporcionarão mais pontos de acesso do que um conjunto de descritores atribuídos pelo indexador, do mesmo modo, com certeza, que o texto integral do documento. É admissível, portanto, que as bases de dados de texto livre possibilitarão, em geral, maior revocação do que aquelas que se apóiam na indexação feita por seres humanos.

Outro fator importante que influi no desempenho de um sistema de recuperação é a especificidade com que se pode descrever o conteúdo temático do documento. Os termos do *UNBIS thesaurus* são bastante específicos ao descrever a maioria dos aspectos do conteúdo temático do item apresentado na ilustração 3. O texto livre, no entanto, proporciona maior especificidade por possibilitar a recuperação a partir dos nomes dos líderes do Oriente Médio, enquanto a indexação permite apenas que se faça uma busca no nível de 'líderes políticos'.

Quanto mais pontos de acesso forem providos para a recuperação, mais alta será a revocação possível, porém, provavelmente, menor será a precisão. Um dos motivos disso é simplesmente o fato de que quanto mais pontos de acesso são providos, maior é a probabilidade de que alguns digam respeito a aspectos bastante secundários do documento. Assim, o consulente que recebesse o item da figura 3 numa busca relativa a Arafat julgaria que ele não teria qualquer utilidade para si porque trata de Arafat de uma forma muito sucinta e tangencial.

Quanto mais pontos de acesso forem providos, maior também será a possibilidade de que venham a ocorrer relações espúrias. Como vimos no capítulo 11, essas relações são de dois tipos: 1) falsas associações, 2) relações incorretas entre termos. Muitas possibilidades são vislumbradas na figura 3. Por exemplo, o resumo expandido causaria a recuperação desse item numa busca sobre entrevistas telefônicas com líderes do Oriente Médio (ou qualquer dos líderes mencionados), e a indexação exaustiva causaria sua recuperação durante uma busca

---

* Lamentavelmente, os resultados desses estudos defeituosos continuam sendo divulgados como se tivessem validade (ver, por exemplo, Olson e Boll, 2001).

sobre líderes políticos dos Estados Unidos. Trata-se de falsas associações porque os termos que causaram a recuperação não têm, essencialmente, relação entre si no documento (ENTREVISTAS TELEFÔNICAS não se relaciona diretamente com LÍDERES nem ESTADOS UNIDOS se relaciona diretamente com LÍDERES POLÍTICOS).

Um tipo mais sutil de relação espúria acha-se exemplificado na indexação seletiva ou mesmo no título. Qualquer um dos dois causará a recuperação desse item durante uma busca sobre atitudes do Oriente Médio em face dos Estados Unidos. Neste caso, os termos ORIENTE MÉDIO, ATITUDES e ESTADOS UNIDOS têm relação direta entre si, mas a relação é ambígua.

Quanto mais extenso o registro, maior a chance de que venham a ocorrer relações espúrias. Estas, evidentemente, causam menor precisão.

Outras lições sobre diferenças entre vocabulário controlado e linguagem natural podem ser tiradas da figura 103. Neste caso, o *UNBIS thesaurus* alcança um resultado medíocre na indexação do item. O resumo é bem mais específico do que os termos controlados: o tesauro não possui termo para 'alunos monitores' [*peer tutoring*] ou mesmo 'monitoria' [*tutoring*]. Esse exemplo também ilustra o fato de a linguagem natural tender a ser mais redundante do que os termos controlados de indexação. Por exemplo, o resumo contém o termo *programmed learning* [ensino programado] e *programmed instruction* [instrução programada], de modo que esse item seria recuperado por qualquer um desses termos que a pessoa quisesse eventualmente usar. É provável que o texto integral de um documento proporcione notável redundância, aumentando as chances de vir a incluir uma expressão empregada por quem faz a busca, assim melhorando a revocação.

A indexação feita por seres humanos é, naturalmente, um processo intelectual subjetivo, e os indexadores nem sempre incluem um assunto que deveria ser incluído, representam um assunto com o melhor termo possível ou explicitam alguma relação de interesse potencial para certos usuários. A completeza e redundância do texto completo evita esse tipo de problema. Horty (1962), o verdadeiro pioneiro das buscas em texto completo, reconheceu isso há mais de 20 anos:

> Quando o texto completo dos documentos é utilizado como base de um sistema de recuperação, as consultas não ficam amarradas à maneira como os documentos foram indexados. Quase inevitavelmente o indexador desconhece certos assuntos aos quais a cláusula jurídica é aplicável ou seria aplicável no futuro. Todavia a pesquisa, por sua própria natureza, determina que o pesquisador procure relações inéditas entre vários assuntos; relações que podem não ter sido antecipadas pelo indexador. Ao dispensar por completo o índice e recorrer ao texto original em cada busca, essas novas relações podem ser encontradas (p. 59).

Por outro lado, naturalmente, esta própria redundância cria grandes problemas quando os textos de muitos documentos são combinados para formar uma grande base de dados — há muitas formas pelas quais um assunto pode ser expresso num texto completo e, em alguns casos, o assunto é representado implícita e não explicitamente (O'Connor, 1965), dificultando uma revocação alta. Um vocabulário controlado reduz a diversidade da terminologia. Além disso, ao ligar semanticamente termos que tenham relação entre si, ajuda o usuário a identificar todos os termos que seriam necessários para realizar uma busca completa.

Outro fator a ser levado em conta é a 'recentidade'. Novos termos surgirão nos títulos ou resumos muito antes de surgirem num vocabulário controlado. Para novos assuntos, portanto, a linguagem natural provavelmente vence sem esforço. A precisão será melhor porque o vocabulário controlado não possibilitará uma busca específica. É provável também que a revocação seja melhor porque quem faz a busca não terá de adivinhar quais os termos a serem empregados. Finalmente, o uso do vocabulário controlado costuma ser preferido pelo especialista em informação, que domina inteiramente as diretrizes e regras que o respaldam, enquanto a linguagem natural conta com a preferência do usuário especialista num assunto. Deschâtelets (1986) é um autor que chamou atenção para a importância de se fazer com que a linguagem controlada se aproxime tanto quanto possível da linguagem natural da respectiva área.

Até agora identificamos diversas características do texto livre e da indexação com termos controlados feita por seres humanos, tendo relacionado ambos com

THE USE OF PEER TUTORING AND PROGRAMMED RADIO INSTRUCTION; VIABLE ALTERNATIVES IN EDUCATION

Hannum, W. H..; Morgan, R. M.
1974, 38p.

Florida State University
College of Education
Center for Educational Technology
Tallahassee, Florida 32306

Educational radio*
Programmed instruction*
Developing countries
Nonformal education
Teachers

Educators in developing countries are likely to achieve more by applying the principles rather than the things of educational technology. The principles of program learning have been shown to be effective in promoting learning in a wide variety of circumstances. The most effective instructional materials can be developed through use of the principles of programmed instruction and mastery learning. Radio, when combined with the use of peer tutors, can be an effective educational tool in developing countries. The concepts of programmed learning and mastery learning can be incorporated in the design of educational radio programs. Such programs, accompanied by peer tutors, can accomplish the total educational effort within the resources of many developing countries. This type of educational system is a viable alternative to traditional formal educational. Such a system should be tried in several developing countries to explore its full potential.

FIGURA 103

Comparação entre resumo e indexação com vocabulário controlado

O resumo foi reproduzido de *A.I.D. Research & Development Abstracts* com permissão do Center for Development Information and Evaluation, United States Agency for International Development
Os termos assinalados com asterisco * são aqueles que o indexador considerou como os mais importantes para este item

seus prováveis efeitos sobre a revocação e a precisão. Essas relações são resumidas na figura 104. Isso deixa evidente que a situação é complexa, uma vez que alguns fatores favorecem os termos controlados e outros favorecem o texto livre. A especificidade das palavras do texto costuma melhorar a precisão, mas dificultar a obtenção de revocação alta, pelo menos durante buscas 'conceituais' genéricas, enquanto a extensão do texto costuma melhorar a revocação, porém diminuir a precisão. A preferência por um ou por outro numa situação determinada será bastante influenciada pelo tipo de busca a ser feita: uma busca conceitual genérica recomendará os termos controlados, uma busca altamente específica (principalmente uma que envolva nomes de pessoas, organizações, etc.) recomendará o texto livre, uma busca realmente exaustiva sobre um assunto (por exemplo, toda referência possível a algum medicamento) recomendará o texto completo, enquanto uma busca altamente seletiva (somente os itens mais importantes) provavelmente recomendará a indexação com termos controlados.

Em geral, outros autores chegaram a conclusões semelhantes. Por exemplo, Fugmann (1985) salienta que as buscas em linguagem natural produzem bons resultados no caso de 'conceitos particulares', mas não de 'conceitos genéricos'; Dubois (1987) afirma que uma das vantagens do texto livre é que 'não ocorre demora na incorporação de novos termos'; e Perez (1982) afirma que "um vocabulário controlado pode resultar em perda de precisão" enquanto o texto livre "não perde a especificidade". Knapp (1982) menciona 'assuntos específicos', 'assuntos quentes' e 'nova terminologia' como exemplos de casos em que a linguagem natural provavelmente será mais útil.

Os aspectos relativos aos custos devem, naturalmente, ser também levados em conta numa comparação entre linguagem natural e vocabulário controlado. O custo do processamento intelectual por seres humanos continua a subir rapidamente em relação ao custo do processamento por computador, e a indexação que utiliza vocabulário controlado exige mão-de-obra intensiva e cara. A construção e manutenção de um vocabulário controlado podem também custar caro. Na medida em que aumenta a disponibilidade de textos, a baixo custo, em formato eletrônico, como subproduto de atividades editoriais ou de disseminação de informações, é natural que os administradores de serviços de informação analisem cautelosamente a situação, a fim de decidir se as vantagens da indexação com vocabulário controlado realmente justificam os custos adicionais.

Do ponto de vista da relação custo–eficácia pode-se ver essa comparação como uma negociação entre entrada e saída. Ao abandonar a indexação feita por seres humanos e os vocabulários controlados, é bem provável que haja uma redução dos custos na entrada. No entanto, alcança-se essa redução às expensas de custos mais elevados na saída, pois se lança uma sobrecarga intelectual maior sobre os ombros do usuário da base de dados. Entre os fatores que influem na decisão em torno dessa negociação entre entrada e saída estão a quantidade de documentos e buscas envolvidas, os custos do indexador e da pessoa que faz as buscas, bem como o grau de importância atribuível aos resultados de uma busca.

Fatores que favorecem a revocação	***Efeito do tipo de representação***
Extensão do registro (número de pontos de acesso)	A maioria das representações em texto livre (com exceção apenas dos títulos) será mais longa do que um conjunto de termos de indexação atribuídos. Isso costuma melhorar a revocação, mas reduz a precisão (aumentam os casos de 'leve menção' e relações espúrias).
Redundância	Será comumente maior em texto livre, aumentando as chances de se encontrar um item específico. No entanto, a grande variedade de modos como um assunto se acha representado numa grande base de dados formada de textos dificulta a obtenção de revocação alta.
Presença de termos 'conceituais' genéricos	Muito mais provável de ocorrer numa representação baseada em vocabulário controlado. Pode estar implícita e não explícita no texto.
Ligação de termos semanticamente relacionados	Favorece nitidamente o vocabulário controlado bem-construído.
Fatores que favorecem a precisão	
Especificidade	O texto livre comumente será mais específico, favorecendo a precisão. A diversidade dos modos como os conceitos são representados dificulta muito, no entanto, a obtenção de uma revocação alta em buscas 'conceituais' genéricas. Em buscas deste tipo serão preferíveis os termos controlados relativamente mais genéricos.
Fatores que influem em ambos	
Atualidade	As representações em texto livre serão sempre mais atuais. Para encontrar um assunto novo em folha num sistema baseado em vocabulário controlado, o usuário talvez tenha de experimentar com diversos termos (diminuindo a precisão), e mesmo assim talvez não consiga encontrar tudo sobre o assunto (diminuindo a revocação).
Hábito	Os especialistas em informação totalmente habituados com um vocabulário controlado, conseguirão utilizá-lo de maneira mais eficaz do que outros o fariam. O 'usuário final' pode sentir-se melhor com a linguagem natural que ocorre em documentos de sua área de conhecimento.

FIGURA 104
Os prós e contras do texto livre *versus* vocabulário controlado

Revisão de estudos afins: antes de 1980

Os primeiros trabalhos escritos sobre a experiência com buscas em textos completos, no campo jurídico, estavam impregnados de um grande entusiasmo

em face desse novo recurso. Seus autores, no entanto, não realizaram experiência alguma visando a comparar as buscas em texto completo com a indexação com vocabulário controlado.

Swanson (1960) formou uma pequena coleção de teste, contendo 100 artigos sobre física nuclear, e determinou quais deles eram relevantes para cada uma de um total de 50 questões. A coleção foi também indexada com cabeçalhos de assuntos 'criados especialmente para o campo da física nuclear'. As buscas em textos completos, que contaram com a ajuda de uma 'coleção de grupos de palavras e expressões, a modo de tesauro', apresentaram, segundo Swanson, resultados superiores àqueles logrados pelas buscas em cabeçalhos de assuntos.

A pesquisa 'Cranfield' sobre as características e o desempenho de linguagens de indexação parece ter tido profunda influência no convencimento de muitos profissionais da informação quanto às vantagens da linguagem natural na recuperação da informação. Segundo foi relatado por Cleverdon et al. (1966), tratava-se de um estudo experimental controlado. Uma coleção de teste contendo 1 400 trabalhos de pesquisa, a maioria dos quais relacionada à aerodinâmica, foi indexada de três formas diferentes: 1) os conceitos estudados eram anotados (por exemplo, 'perdas em cascata'), 2) os conceitos eram decompostos nas palavras que os constituíam, no singular ('cascata', 'perda'), e 3) os conceitos afins eram reunidos para formar 'ligações'ou 'temas' (por exemplo, compressor de fluxo axial/perda em cascata). Os itens foram indexados exaustivamente: não era raro haver de 30 a 50 'conceitos' por item.

Compilou-se um grupo de 221 questões de teste. Essas questões foram elaboradas por especialistas e se baseavam em trabalhos de pesquisa reais dos quais eles mesmos eram autores. A coleção de teste foi examinada minuciosamente por estudantes de pós-graduação do College of Aeronautics (em Cranfield, Inglaterra), e os itens que apresentassem qualquer 'relevância' imaginável eram enviados à pessoa que propusera a questão, a fim de que fossem avaliados de acordo com uma escala de cinco pontos (dos quais uma categoria correspondia a 'absolutamente não-relevantes'). Como resultado, ficava-se sabendo quais os itens da coleção que eram relevantes para cada questão do teste (pelo menos aos olhos do autor da questão) e quais não eram.

Todo o estudo foi realizado como uma espécie de simulação. Foram 'montados' diferentes tipos de vocabulários, variando do mais elementar (palavras simples sem quaisquer controles, fusão de singular/plural, controle da forma das palavras [busca em radicais de palavras], controle simples de sinônimos) ao mais complexo (agrupamento dos termos em hierarquias como apareceriam numa verdadeira classificação hierárquica). Cada questão foi proposta 33 vezes à coleção de teste, cada uma delas correspondendo ao teste de um vocabulário diferente (33 ao todo), possibilitando uma comparação dos resultados alcançados pelos vários vocabulários. Ao serem combinadas as medidas de revocação e precisão numa única medida de desempenho (denominada 'revocação normalizada'), os vocabulários constituídos de termos de uma única palavra da linguagem natural (com controle das formas dos vocábulos, com controle de sinônimos ou sem qualquer controle) superaram em desempenho todos os outros.

O estudo Cranfield gerou muita polêmica e suscitou muitas críticas ao longo dos anos. Grande parte dessas críticas, no entanto, teve origem numa falta de compreensão daquilo que o estudo realmente realizou. Por exemplo, Soergel (1985) sugeriu que tanto a indexação quanto a linguagem de indexação eram de qualidade duvidosa. Como eu fui um dos indexadores, posso testemunhar os grandes cuidados que cercaram a indexação — cuidados muito maiores do que os que provavelmente ocorrem num ambiente de produção habitual — e que os indexadores possuíam experiência anterior bastante respeitável. Mesmo hoje em dia ainda surgem críticas. Alguns autores procuraram desacreditar os resultados de Cranfield com o argumento de que, como as questões se baseavam em documentos reais, isso criaria um viés favorável à linguagem natural. É difícil entender essa crítica, uma vez que os itens considerados relevantes pelos autores das questões não eram os itens nos quais haviam baseado as questões.

De qualquer modo, não é minha intenção defender aqui os estudos Cranfield, mas meramente salientar que, defeituosos ou não, levaram muitas pessoas a acreditar que, pelo menos em certas circunstâncias, os sistemas baseados em linguagem natural teriam um desempenho tão bom ou melhor do que os baseados em vocabulários controlados. Em alguns dos trabalhos que escreveu depois da realização dos testes de Cranfield, Cleverdon sugeriu que um sistema com base na linguagem natural, se implementado de forma apropriada, sempre teria desempenho superior ao de um sistema baseado em controle de vocabulário. Pouco mais tarde, Klingbiel (1970) valeu-se dos resultados de Cranfield, aliados à sua própria experiência no Defense Documentation Center, para afirmar que "vocabulários controlados altamente estruturados são obsoletos para a indexação e a recuperação" e que "a linguagem natural da prosa científica é totalmente adequada à indexação e à recuperação". Pouco depois, Bhattacharyya (1974) diria:

> Os resultados de vários experimentos visando a testar e avaliar as linguagens de indexação, realizados durante a última década, demonstraram, repetidas vezes, a força da linguagem natural, com o mínimo ou sem nenhum controle, como a melhor linguagem de indexação (isto é, levando em conta tanto a eficácia quanto a eficiência de recuperação) (p. 235).

Após os estudos em Cranfield, e influenciados por eles, diversos pesquisadores chegaram a conclusões semelhantes quanto aos méritos da linguagem natural na recuperação da informação. Por exemplo, Aitchison et al. (1969–1970) empreenderam alguns testes visando a obter elementos que ajudassem nas decisões acerca da indexação da base de dados INSPEC. Compararam-se os resultados de buscas feitas em: 1) título, 2) títulos mais resumos, 3) termos de indexação utilizados na publicação impressa *Science Abstracts*, 4) indexação feita por seres humanos com 'linguagem livre', e 5) termos controlados extraídos de um rascunho de tesauro compilado pelo pessoal do INSPEC. O ambiente do

teste consistia em 542 artigos no campo da eletrônica e 97 questões formuladas por pesquisadores. Fizeram-se avaliações no sentido de determinar quais artigos eram relevantes para quais questões. Tomou-se o cuidado de estabelecer algum nível de 'equivalência' entre as estratégias adotadas nas várias modalidades de busca. Verificou-se que a recuperação baseada no rascunho de tesauro proporcionou resultados melhores do que qualquer uma das outras modalidades de busca. Todavia, recomendava-se que a atribuição pelos indexadores de termos da linguagem livre, que haviam ficado em segundo lugar quanto ao desempenho, deveria ser o método adotado. A base de dados INSPEC incorporou posteriormente tanto termos de tesauro quanto termos de texto livre.

Em importante estudo, Keen e Digger (1972) compararam o desempenho de vários tipos de vocabulários no campo da ciência da informação. As principais características desse teste podem ser assim resumidas:

1. Foram utilizadas cinco linguagens de indexação diferentes: UL, uma linguagem pós-coordenada, não-controlada, construída por indexadores mediante a seleção de palavras dos próprios documentos; CT, uma linguagem pós-coordenada de 'termos comprimidos', contendo menos de 300 termos, com estrutura de tesauro; Pre-HS, uma linguagem pré-coordenada, hierarquicamente estruturada, na forma de um esquema de classificação facetada; HS, uma linguagem hierarquicamente estruturada (o esquema de classificação é modificado de modo a permitir que seja utilizado de maneira pós-coordenada); Pre-RI, uma linguagem pré-coordenada na qual os termos da classificação hierárquica são combinados em locuções de indexação ('analetos') com o emprego dos operadores relacionais de Farradane.
2. Uma coleção de teste de 800 documentos sobre biblioteconomia e ciência da informação foi indexada pelos dois pesquisadores, que utilizaram cada um dos cinco vocabulários.
3. Os índices criados eram inteiramente manuais, sendo que o índice pós-coordenado foi montado com fichas de coincidência óptica.
4. Sessenta e três pedidos de buscas, obtidos junto a bibliotecários e outros especialistas em informação, foram processados nesses índices.
5. As buscas foram realizadas por 19 estudantes de biblioteconomia e ciência da informação, que empregaram um plano experimental de quadrado latino.
6. Vinte auxiliares de ensino da área elaboraram julgamentos de relevância dos pedidos do teste em relação a cada documento da coleção.
7. Os testes foram realizados com diferentes 'versões' dos cinco índices. Essas versões refletiam mudanças introduzidas na linguagem de indexação ou na política de indexação. As principais variáveis assim examinadas foram o efeito da exaustividade da indexação (isto é, o número de termos atribuídos por documento), a especificidade do vocabulário, diferentes métodos de coordenação dos termos no momento da busca, o grau em que os termos são interligados (por remissivas ou estrutura hierárquica) num vocabulário; união de termos afins no momento da indexação (isto é, 'compartimentagem'), o emprego dos operadores relacionais e a provisão de 'contexto' no arquivo de buscas (quem realiza a busca num índice de coincidência óptica é remetido, pelo número do documento, a um 'arquivo de contextos' onde uma entrada de índice alfabético em cadeia representa o conteúdo temático específico estudado no documento, o que equivale aproximadamente ao contexto provido num índice pré-coordenado).

As diferentes linguagens foram empregadas em diferentes comparações (quer dizer, nem todas as comparações são relevantes para todas as linguagens), e se utilizou em algumas dessas comparações um subconjunto de 241 documentos e 60 pedidos de buscas. Os resultados das diversas comparações são apresentados, em sua maioria, sob a forma de coeficientes de revocação e números absolutos de itens não-relevantes recuperados.

Talvez de maneira não muito imprevisível, essa pesquisa produziu resultados que tendem a corroborar os resultados de estudos anteriores:

> As linguagens não-controladas testadas tiveram, em geral, um desempenho tão bom quanto o das linguagens controladas, ao proporcionar uma eficácia de recuperação coerentemente boa e um desempenho de eficiência que jamais chegou a ser tão ruim quanto o da pior linguagem controlada, nem tão bom quanto as melhores, e em nenhum caso as diferenças foram estatisticamente importantes (volume 1, p. 166-167).

Os pesquisadores, além disso, afirmam que:

> o que se prescreve para a melhor linguagem de indexação é, evidentemente, que apresente a mais alta especificidade possível sem ter de empregar dispositivos de precisão que sejam mais complexos do que a simples coordenação (e com pouca ou nenhuma pré-coordenação de termos). E, realmente, parece que as palavras simples da linguagem natural inglesa se aproximam da provisão desse nível ótimo de especificidade (volume 1, p. 169).

Keen e Digger chegaram a sugerir que agora estava bem comprovado o argumento contra os vocabulários controlados, ao ponto de afirmarem que "esta deve ser a última vez em que as tradicionais linguagens controladas de indexação são humilhadas, por ter ficado demonstrado que não oferecem vantagem alguma" (volume 1, p. 170).

Lancaster et al. (1972) realizaram um estudo sobre buscas em linha feitas por pesquisadores da área biomédica no Epilepsy Abstracts Retrieval System (EARS). Seu objetivo era determinar a eficiência com que esses pesquisadores podiam fazer buscas no texto de resumos no campo da epilepsia, tendo sido efetuadas algumas comparações entre texto livre e termos controlados. Constatou-se que, em 47 buscas, o emprego dos termos de indexação atribuídos pela *Excerpta Medica* proporcionou cerca de metade da revocação que fora obtida com as buscas feitas em resumos. Observe-se, contudo, que os resumos geralmente ofereciam muito mais pontos de acesso, de modo que a comparação foi

mais a respeito da extensão do registro do que uma verdadeira comparação entre buscas em texto livre *versus* buscas com termos controlados. Os pesquisadores concluíram que a busca em texto resultou em melhor revocação devido a: a) número de pontos de acesso, b) maior redundância, c) maior coincidência entre os termos empregados pelos usuários e as palavras do texto, d) erros e incoerências na indexação feita por seres humanos, e e) termos de indexação coincidentes entre si. Levantou-se a hipótese de que o desempenho melhoraria grandemente se algum tipo de 'tesauro de busca' fosse acrescentado ao sistema.

Utilizando documentos e questões reunidos por Lancaster na avaliação que este efetuou sobre o sistema MEDLARS (Lancaster, 1968a), Salton (1972) apresentou resultados que sugeriam que seu sistema SMART superaria em desempenho as dispendiosas atividades de indexação e controle de vocabulário associadas ao MEDLARS. Essa comparação difere um pouco da comparação convencional de buscas em bases de dados que empregam linguagem natural e que empregam vocabulários controlados. O SMART não funciona com base na álgebra booleana, mas por meio de uma espécie de 'coincidência de padrões', em que os textos dos resumos são cotejados com os textos dos pedidos feitos em linguagem natural, e o usuário recebe um conjunto de opções de busca com variados níveis de complexidade. Nos estudos de Salton, o SMART parecia superar o desempenho do MEDLARS somente quando se aplicava uma retroalimentação proporcionada pelo usuário. Ou seja, os usuários avaliavam os resultados preliminares da busca e esta era repetida com base na retroalimentação fornecida pelo usuário a respeito da relevância. Isso suscita a questão de saber qual seria o desempenho do MEDLARS se adotasse a retroalimentação de relevância. O SMART voltará a ser examinado no próximo capítulo.

Importante estudo realizado nesse período é freqüentemente esquecido. Cleverdon (1977) comparou buscas em linguagem natural e com termos controlados num subconjunto da base de dados da NASA formado por 44 000 itens. Foram realizadas buscas em linha em quatro centros em cada um dos quais se faziam dez buscas. Cada busca era feita de um modo por uma pessoa e de um modo diferente por uma segunda pessoa. As duas pessoas que realizavam a busca sobre o mesmo assunto, cada uma de um modo diferente, primeiramente analisavam a solicitação, a fim de chegar a um acordo sobre aquilo que o consulente desejava. Essas modalidades de busca eram: a) somente em termos controlados, b) linguagem natural dos títulos e resumos, c) termos controlados combinados com linguagem natural, e d) linguagem natural com o auxílio de uma lista de 'conceitos associados'. Constatou-se que as buscas em linguagem natural resultaram numa revocação acentuadamente mais alta e pouco diferiam, quanto à precisão, das buscas com termos controlados. A conclusão de Cleverdon, corretamente, foi que a extensão do resumo fora a principal causa disso.

Infelizmente, o estudo de Cleverdon é prejudicado pelas deficiências formais com que foi descrito. Por exemplo, as buscas em que tanto os termos controlados

quanto a linguagem natural foram utilizados tiveram um desempenho bem inferior, tanto quanto à revocação quanto à precisão, do que as buscas que envolviam apenas linguagem natural. Isso é exatamente o oposto do que seria natural, sendo difícil de explicar, principalmente porque essas buscas de 'modo conjunto' recuperaram duas vezes mais itens do que as buscas em linguagem natural. Cleverdon não apresenta explicação para essa anomalia. Outra anomalia é que as buscas em linguagem natural que contaram com a ajuda do 'arquivo de conceitos associados' também tiveram desempenho bem inferior ao das buscas realizadas com o emprego somente da linguagem natural. Isso também não é explicado com clareza, ficando difícil para os leitores do relatório de Cleverdon chegar a suas próprias conclusões, uma vez que o próprio 'arquivo de conceitos associados' não é descrito completamente. Tudo que se pode conjeturar a partir da descrição de Cleverdon é que esse arquivo resultou da co-ocorrência de termos nos títulos de documentos da coleção.

Trabalho posterior de Martin (1980) oferece alguns esclarecimentos, mas, por sua conta, aumenta ainda mais o mistério. Ele esclarece que o componente de linguagem natural da base de dados consistia em palavras simples extraídas por computador dos títulos e resumos e posteriormente revistas por seres humanos a fim de eliminar 'palavras proibidas' e normalizar o vocabulário com a exclusão de grafias e formas lexicais variantes. O arquivo de conceitos associados era um arquivo das palavras-chave extraídas apenas dos títulos e que mostrava, para cada um deles, as palavras-chave que ocorriam com maior freqüência nos títulos. Martin resume os resultados assim:

	Revocação (%)	*Precisão (%)*
Termos controlados	56	74
Linguagem natural	78	63
Linguagem natural mais termos controlados	71	45

E, então, acrescenta que "para cada documento relevante recuperado pela linguagem controlada, a linguagem natural sozinha recuperou 1,4, a linguagem natural mais a linguagem controlada 1,6 [...]", o que é totalmente incompatível com os valores de revocação/precisão apresentados. Martin também esclarece que as buscas em 'linguagem natural mais termo controlado' incluem algumas que envolviam somente termos controlados (onde a pessoa que fazia a busca não viu necessidade de adicionar a linguagem natural) e, portanto, "elas não representavam todo o potencial de LC [linguagem controlada] mais LN [linguagem natural]". As incoerências nos resultados e nas afirmações sobre eles, bem como preocupações acerca das instruções passadas às pessoas que faziam as buscas, lançam alguma dúvida sobre a validade dessa comparação.

Somente um estudo realizado durante esse período afirma ter encontrado resultados superiores para a indexação feita por seres humanos com o emprego de um vocabulário controlado. Hersey et al. (1971) utilizaram um subconjunto da base de dados do Smithsonian Science Information Exchange (SIE, formada

por 4 655 descrições de projetos, na comparação que fizeram entre texto livre e 'indexação por cientistas'. A indexação envolveu o emprego de códigos de assuntos, atribuídos por especialistas, e extraídos de um esquema de classificação desenvolvido especialmente para esse fim. Para 27 buscas realizadas no próprio SIE, foram alcançados os seguintes resultados:

	Revocação (%)	*Precisão (%)*
Texto de descrições de projetos	66	81
Indexação de assuntos	95	95

Mais uma vez as deficiências do relato dificultam para o leitor a compreensão exata do que foi feito. As questões utilizadas foram umas que haviam sido 'anteriormente formuladas', mas não se esclarece se os resultados concernentes à indexação de assuntos foram obtidos quando as buscas foram originalmente feitas para os usuários, ou se foram obtidos mais tarde, no momento da realização das buscas em texto livre. Os seguintes pontos também são obscuros: de que modo foram feitas as avaliações de relevância (aparentemente foram feitas pelo pessoal do SIE e não pelos solicitantes originais), de que forma o pedido foi entregue a quem fazia a busca em texto livre, e se foram ou não impostos controles às pessoas que faziam as buscas, a fim de se conseguir certo nível de equivalência de método entre busca em texto e busca em termos de indexação.

Cada um desses fatores teria grande influência nos resultados do estudo. Por exemplo, se o pedido usado como base para a busca em texto livre não estivesse nas palavras originais do solicitante, mas houvesse sido 'negociado' mediante interação com o pessoal do SIE, os resultados da comparação poderiam muito bem apresentar um viés para os códigos de assuntos. O fato de os resultados desse estudo relativos à revocação/precisão terem sido bem mais altos do que os de outras pesquisas, e muito mais altos do que os obtidos durante o funcionamento rotineiro de sistemas de recuperação (Lancaster, 1968a), aliado a um relato muito impreciso, suscita sérias dúvidas quanto à validade dessa comparação. Os valores de precisão excepcionalmente altos explicam-se, porém, parcialmente, pelo fato de que o arquivo de teste de descrições de projetos era realmente uma fusão de quatro arquivos de teste separados sobre áreas temáticas completamente diferentes.

Um estudo suplementar nessa base de dados foi realizado pelo Biological Sciences Communication Project, da George Washington University, utilizando 12 questões do SIE. As buscas nos códigos de assuntos recuperaram 91 projetos, 74 dos quais foram considerados relevantes (precisão de 81%), enquanto as buscas em texto recuperaram 70, dos quais 43 foram considerados relevantes (precisão de 61%). Combinando-se os resultados das buscas em texto livre e com códigos de assuntos, conclui-se que a revocação foi de cerca de 50% para texto e 90% para indexação de assuntos, porém alguns itens só foram recuperados por cada uma das modalidades de busca.

Byrne (1975) utilizou 50 perfis de DSI na base de dados COMPENDEX e comparou os resultados quando as buscas foram realizadas nos títulos, resumos e cabeçalhos de assuntos, além de várias combinações destes elementos. Os resultados de uma modalidade de busca foram comparados com os resultados combinados de todas as modalidades. Empregando este padrão, os cabeçalhos de assuntos sozinhos recuperaram 21% dos itens, os resumos sozinhos 61%, os títulos mais os resumos 75%, e os títulos mais os termos de assuntos 41%. Não é de estranhar, portanto, que, aparentemente, as representações mais extensas tenham resultado em muito melhor revocação. No entanto, não se fizeram avaliações reais de relevância nesse estudo: tudo que foi recuperado foi considerado *ipso facto* como sendo uma resposta apropriada.

Revisão de estudos afins: a partir de 1980

Em geral, as comparações entre buscas em texto livre e com vocabulário controlado realizadas nas décadas de 1960 e 1970 mostravam que o texto livre funcionava tão bem quanto os termos controlados, senão melhor. Tais estudos, porém, foram realizados em arquivos muito pequenos, e às vezes insignificantemente pequenos. Em sua maioria, tratava-se de estudos experimentais, ao invés de envolver serviços de informação reais funcionando em condições de trabalho concretas. A partir de 1980 alguns estudos foram realizados com bases de dados de maior porte e/ou envolvendo serviços verdadeiramente operacionais.

Markey et al. (1980) empreenderam uma análise de enunciados de busca com vocabulário controlado e texto livre em buscas em linha na base de dados ERIC. Também realizaram 'testes de buscas em linha', comparando vocabulário livre e controlado, mas usando somente seis assuntos. Concluíram que o texto livre resultou em revocação mais alta e os termos controlados resultaram em mais alta precisão. Igual a muitos outros estudos, o relato do teste é lamentavelmente inadequado. Não há informações sobre como foram feitas as avaliações de relevância nem sobre como as buscas foram realizadas, de modo que o leitor não sabe se foi feito algum esforço para 'controlar' as estratégias de busca, a fim de evitar o favorecimento de uma das modalidades de busca. Os escores insolitamente elevados (93% de revocação e 71% de precisão para texto livre, e 76% de revocação e 95% de precisão para termos controlados) lançam dúvida sobre a validade desse estudo.

Diversos estudos foram realizados no campo do direito. Coco (1984) utilizou uma base de dados sobre casos em tribunais itinerantes (1960–1969) e 50 'problemas de pesquisa' verdadeiros extraídos de um estudo de 1977 do Federal Judicial Center, a fim de comparar a recuperação nos sistemas WESTLAW e LEXIS. O LEXIS inclui somente o texto dos pareceres vinculados a esses casos, enquanto o WESTLAW acrescenta 'componentes editoriais' ao texto dos pareceres, inclusive várias formas de sinopses. O objetivo declarado desse estudo era comparar os resultados de buscas baseadas somente no texto com os alcançados com o texto mais acréscimos editoriais. Como as buscas no WESTLAW foram executadas com

e sem os acréscimos editoriais, a comparação com o LEXIS tornou-se totalmente desnecessária e só serviu para confundir o leitor. De qualquer modo, a comparação entre LEXIS e WESTLAW não poderia ser considerada inteiramente válida porque as bases de dados não eram exatamente comparáveis. Como diz Coco, "os sistemas continham *aproximadamente* [grifo meu] o mesmo número de casos para esse período". Além disso, não houve qualquer esforço sistemático para determinar se os casos recuperados eram ou não de alguma forma relevantes para os problemas de pesquisa.

Se o único exemplo apresentado por Coco for representativo de todos os itens da base de dados, o texto ampliado do WESTLAW é quase duas vezes o tamanho do texto do parecer sozinho. Não é de estranhar, portanto, que tenha recuperado mais casos (913 contra 728, embora não se saiba quantos mais eram 'relevantes'). De fato, seria razoável supor que o dobro da extensão de texto causaria um aumento superior a 20% do número de casos recuperados. O fato de isso não ter ocorrido deve ser em parte atribuível à coincidência de termos entre o texto e os acréscimos editoriais. Os resultados do estudo eram totalmente previsíveis desde o início, e seria dispensável esse tipo de pesquisa para nos dizer que dobrando o tamanho do texto crescerá o número de itens recuperados.

Blair e Maron (1985) realizaram um estudo bastante extenso sobre uma base de dados jurídicos, que utilizava o sistema STAIRS (cerca de 350 000 páginas de texto, ou 40 000 documentos, e 40 pedidos de informação). Auxiliares advocatícios realizaram buscas exaustivas, iterativas, em linha, e só as interromperam quando os advogados para quem trabalhavam se consideraram satisfeitos, pois pelo menos 75% das referências relevantes haviam sido recuperadas. Por amostragem, no entanto, os pesquisadores calcularam que se alcançara não mais de 20% de revocação. Concluem que os resultados de seu estudo lançam séria dúvida sobre a eficácia das buscas em texto completo e, com base em algumas análises de custos muito duvidosas, que as buscas em texto completo são muito mais caras do que os métodos alternativos. Esquecem por completo o fato de que grandes sistemas que empregam vocabulário controlado talvez não alcancem um desempenho melhor. Por exemplo, um estudo de 535 buscas no MEDLINE, realizadas por 191 pessoas diferentes, mostrou que elas apresentaram uma revocação média de apenas 23% e uma precisão de 67% (Wanger et al., 1980). Dabney (1986a), embora tomando por base em grande parte os resultados de Blair e Maron, oferece um excelente estudo dos problemas da recuperação em texto completo no campo jurídico. Respostas de McDermott (1986) e de Runde e Lindberg (1986) a Dabney, bem como um comentário com aditamentos de Dabney (1986b), também merecem ser examinados. Salton (1986) produziu uma minuciosa revisão do estudo de Blair e Maron. Ele discorda enfaticamente da conclusão deles segundo a qual bases de dados indexadas por seres humanos provavelmente terão melhor desempenho do que as buscas em textos.

Um dos melhores estudos em que se compara o texto completo com resumos

e indexação controlada foi empreendido por Tenopir (1984). Utilizando a *Harvard Business Review* em linha, Tenopir obteve os seguintes resultados, divididos proporcionalmente entre 31 buscas:

	Texto completo	*Resumos*	*Termos controlados*
Número de documentos recuperados (média)	17,8	2,4	3,1
Documentos relevantes recuperados (média)	3,5	1,0	1,2
Revocação (relativa à fusão de todos os métodos)	73,9	19,3	28,0
Precisão	18,0	35,6	34,0
Custo por busca (em US$)	20,57	4,95	5,32
Custo por item relevante recuperado (em US$)	7,86	3,89	3,54

As cifras de Tenopir relativas a custos não podem ser levadas muito a sério, pois ela incluiu os custos da aquisição de cópias completas dos documentos para a realização dos julgamentos de relevância, enquanto na vida real isso raramente aconteceria (isto é, os usuários fariam seus julgamentos com base nos títulos e/ou resumos mostrados em linha). Talvez o resultado mais importante da pesquisa de Tenopir seja ter verificado que as buscas em termos controlados recuperaram alguns itens que não foram recuperados com texto completo, e vice-versa, demonstrando a necessidade de ambos os métodos.

Posteriormente, Ro (1988) realizou estudo dando seguimento à pesquisa sobre a base de dados da *Harvard Business Review*, o qual produziu resultados semelhantes aos alcançados por Tenopir.

Sievert et al. (1992) descobriu, o que não foi surpresa, que buscas numa base de dados que continha o texto integral de artigos de revistas médicas obtinham melhor revocação do que buscas na base MEDLINE, embora as buscas em texto completo resultassem em muito menor precisão. Em artigo anterior, contudo, chamaram a atenção para os problemas das buscas em texto completo ao analisar os motivos de não-recuperação, na base de dados de textos completos, de itens relevantes recuperados no MEDLINE (Sievert e McKinin, 1989)

Os melhoramentos que a utilização de termos do texto, além dos termos controlados, introduz na revocação foram demonstrados por diversos pesquisadores, inclusive McCain et al. (1987), que compararam os resultados de buscas em cinco bases de dados sobre 11 tópicos das ciências médicas comportamentais.

Vários outros estudos relataram os resultados de buscas em texto completo ou parcial, mas sem fazer comparações com buscas com termos controlados. Alguns desses estudos envolveram sistemas (semelhantes de algum modo ao SMART) que adotam métodos probabilísticos e/ou lingüísticos de ordenação dos documentos, ou parágrafos deles, com base em sua similaridade com enunciados de pedidos ou estratégias de busca. Por exemplo, Bernstein e Williamson (1984) avaliam esses métodos aplicados à Hepatitis Knowledge Base [Base de Conhecimentos sobre Hepatite], e Tong et al. (1985) avaliam técnicas de inteligência artificial aplicadas à recuperação em texto completo numa base de dados de notícias.

Fidel (1992) sugere quais os fatores que favorecerão as buscas com vocabulário controlado e os que favorecerão as buscas em textos. Num estudo de 281 buscas reais efetuadas por 47 especialistas treinados, ela identificou vários fatores que afetam a escolha de termos controlados *versus* palavras do texto feita por quem faz a busca. Ela constatou que existe mais confiança no texto em algumas áreas temáticas do que em outras (embora isso possa estar menos relacionado às características do assunto ou sua linguagem do que à qualidade dos vocabulários controlados usados em várias bases de dados — especialmente sua especificidade — e à qualidade da indexação com vocabulário controlado).

Com a finalidade de melhorar os resultados das buscas, alguns pesquisadores estudaram os efeitos da segmentação de um texto em unidades menores, numa tentativa de melhorar a precisão das buscas sem sérios prejuízos para a revocação. Williams (1998) distingue entre segmentação do discurso (baseado em frases, parágrafos, seções) e segmentação em janelas (divisão do texto em pedaços de tamanho arbitrário). Williams testou a recuperação (coeficientes de revocação e precisão) para parágrafos, páginas, três diferentes janelas (250, 500 e 1 000 palavras) e três janelas superpostas de 250, 500 e 1 000 palavras. A superposição arbitrária foi planejada para evitar a separação de textos afins que, do contrário, ocorreria com a segmentação arbitrária. Williams constatou que a janela superposta de 500 palavras parecia oferecer o melhor resultado global quando medido pela revocação e precisão. Ele conclui que esse tipo de segmentação pode melhorar substancialmente a precisão com uma queda moderada da revocação. Williams refere-se a essa abordagem como 'indexação de passagem de nível'. Não fica claro como essa abordagem representa melhoria em comparação com as buscas por proximidade de palavras, que era empregada em buscas em textos 40 anos antes.

A revisão da literatura aqui incluída concentrou-se em estudos que comparam o desempenho de bases de dados de texto livre com o de bases de dados em que se adota a indexação por meio de vocabulários controlados e quando as buscas são feitas com o emprego de combinações booleanas de termos. Embora outros tipos de estudo tenham sido mencionados, não se procurou fazer uma revisão de toda a literatura sobre buscas em texto e que empregam métodos não-booleanos.

Esta revisão deixa evidente que o imoderado entusiasmo inicial pelas buscas em linguagem natural sofreu um abrandamento com o passar dos anos à medida que se identificavam com maior clareza os problemas que implicava. Alguns dos primeiros estudos baseavam-se em bases de dados de cunho experimental que eram insignificantemente diminutas. Considerando que se pode tolerar uma precisão muito baixa quando se recupera apenas um punhado de itens, é possível conseguir um nível aceitável de revocação. Esta situação se altera substancialmente quando se passa para bases de dados que contêm centenas de milhares de itens. Então, por causa do número de itens recuperados ('sobrecarga de saída'), já não são mais aceitáveis baixos níveis de precisão, sendo analogamente difícil

obter alta revocação com nível aceitável de precisão. Há, porém, indícios (Wanger et al., 1980) de que isso é também verdadeiro no caso de grandes sistemas baseados em vocabulários controlados, não sendo uma peculiaridade exclusiva das buscas em texto livre.

É importante reconhecer a diferença entre as expressões *texto livre* e *texto completo*. As conclusões alcançadas como resultados de estudos sobre bases de dados de texto completo não se transferem automaticamente para bases de dados que contenham algo menor do que o texto completo (por exemplo, resumos). Nas bases de texto completo o problema de escala é agravado. Quer dizer, com uma base de texto completo muito grande será ainda mais difícil alcançar revocação aceitável com precisão tolerável. O texto completo proporcionará maior revocação, porém menor precisão do que uma base de dados que contenha algo menor do que o texto completo. Isto foi claramente demonstrado por Tenopir (1984).

É lamentável que a maioria dos estudos que se propõem a comparar o desempenho na recuperação entre texto livre e um conjunto de termos de indexação selecionados de um vocabulário controlado não cumpra isso. Ao contrário, eles comparam o desempenho na recuperação de registros de extensão variável. Uma comparação válida entre termos controlados *versus* texto livre de per si teria de manter constante a extensão dos registros (por exemplo, todos os tópicos mencionados num resumo teriam de ser traduzidos, até onde fosse possível, para termos controlados equivalentes), bem como a estratégia de busca (isto é, uma estratégia 'conceitual' teria de ser criada e em seguida traduzida exatamente para: a] expressões do texto, e b] termos selecionados do vocabulário controlado). Isso parece que nunca foi feito desde os estudos em Cranfield. Tenopir controlou suas estratégias de busca, mas, como estava utilizando uma base de dados já existente, não pôde controlar a extensão do registro. Conseqüentemente, suas conclusões dizem respeito muito mais à extensão do registro do que à controvérsia sobre linguagem natural/vocabulário controlado.

Também é lamentável o fato de a bibliografia ainda trazer afirmativas disparatadas, baseadas em indícios casuísticos, de defensores de ambos os campos, que se recusam a aceitar o fato de que a linguagem natural e os vocabulários controlados têm ambos suas respectivas vantagens. Para um bom exemplo ver Fugmann (1987).

Um exame meticuloso da bibliografia incluída nesta revisão não me outorga razão alguma para modificar minhas opiniões originais sobre os prós e contras dos dois métodos, conforme se acham resumidas na figura 104. O fato é que cada um deles tem suas vantagens e desvantagens. Os registros em texto livre costumam ser mais extensos e, por isso, proporcionam mais pontos de acesso; freqüentemente incluirão alguns termos mais específicos ou mais atualizados do que aqueles existentes em qualquer vocabulário controlado e, comumente, proporcionarão maior redundância. O vocabulário controlado, por outro lado, impõe coerência na representação do conteúdo temático dos documentos, dispõe

dos termos 'conceituais' genéricos que amiúde não se encontram no texto, e, por meio de uma estrutura hierárquica e remissivas, oferece ao usuário uma ajuda positiva na identificação de termos de busca que sejam apropriados.

Sistemas híbridos

Praticamente todos os autores que escreveram a respeito de buscas em texto livre, inclusive Henzler (1978), Perez (1982) e Muddamalle (1998), bem como a maioria dos autores já citados, chegaram à conclusão, já esperada, de que o sistema de recuperação ideal incluirá uma parte de termos controlados, bem como uma parte de texto livre. São óbvias as vantagens desses sistemas híbridos, descritos e exemplificados há muitos anos por Holst (1966), Uhlmann (1967 e Lancaster (1972). A utilidade do método híbrido é apoiada pelo fato de que, na maioria dos estudos realizados, as buscas em texto livre recuperaram alguns itens relevantes que não foram identificados por buscas com vocabulário controlado, e vice-versa.

O termo *híbrido* é empregado para designar qualquer sistema que funcione com uma combinação de termos controlados e linguagem natural, inclusive aqueles em que ambos os conjuntos de termos são atribuídos por indexadores humanos e aqueles em que uma base de dados pode ser consultada mediante uma combinação de termos controlados atribuídos por seres humanos e palavras que ocorram nos títulos, resumos ou texto completo.

Vejamos, por exemplo, um sistema baseado em três componentes vocabulares independentes:

1. um pequeno vocabulário de códigos de assuntos genéricos, com um total talvez de 300 códigos;
2. uma lista de códigos que representem áreas geográficas; e
3. palavras-chave ou expressões que ocorram nos títulos ou textos dos documentos.

A indexação com esses elementos vocabulares representaria uma economia importante em relação à indexação que empregue um grande vocabulário meticulosamente controlado, por dois motivos:

1. Os códigos de assuntos seriam suficientemente genéricos para serem atribuídos sem muita dificuldade por um indexador que não dispusesse de um alto nível de formação educacional ou especialização num assunto.
2. O número de códigos (temáticos e geográficos) é suficientemente reduzido para que o indexador retenha a maioria deles na memória e dispense a consulta constante a uma lista de um vocabulário.

Embora qualquer um dos elementos do vocabulário, isoladamente, seja relativamente imperfeito, o emprego conjunto de uma palavra-chave (para obter especificidade) e um código temático ou geográfico (para obter o contexto) constitui dispositivo extremamente poderoso. Por exemplo, a palavra-chave *plantas* pode significar algo inteiramente diferente ao ser combinada com um código temático relativo à agricultura ou ao ser combinada com um código semântico relativo à

arquitetura. Igualmente, a palavra-chave *assalto*, associada ao código geográfico relativo ao Iraque, indica uma operação de guerra; por outro lado, quando coordenada com o código geográfico relativo a uma metrópole onde a criminalidade seja alta, é mais provável que signifique roubo. Além disso, o emprego conjunto de códigos de assuntos genéricos, códigos geográficos e palavras-chave é extremamente eficaz para esclarecer relações, mesmo quando essas relações não se acham especificadas explicitamente. Muitas das bases de dados atualmente acessíveis em linha podem ser consultadas com o emprego de combinações de termos controlados e palavras-chave ou expressões que ocorrem nos títulos ou nos resumos, sendo que os últimos permitem maior especificidade.

O vocabulário pós-controlado

Diversos autores salientaram que as buscas em linguagem natural melhoram consideravelmente mediante a elaboração e utilização de várias formas de instrumentos auxiliares de busca. Piternick (1984) descreveu alguns desses instrumentos auxiliares. Deles, o mais evidente seria um 'tesauro de buscas' ou 'vocabulário pós-controlado' imaginado por Lancaster (1972), Lancaster et al. (1972), e, mais detidamente, por Lancaster (1986).

O primeiro sistema desenvolvido para fazer buscas em grandes coleções de textos jurídicos (em Pittsburgh) utilizava uma espécie de tesauro para ajudar no processo de buscas. Tratava-se, simplesmente, de uma compilação de palavras com significados semelhantes, parecendo-se mais com o *Roget's thesaurus* do que com a estrutura de tesauro comumente usado na recuperação da informação. Mesmo sem contar com uma 'estrutura' que se revestisse de alguma importância, esse tesauro era um instrumento auxiliar extremamente útil durante as buscas; como palavras de significado similar são potencialmente substituíveis durante uma busca, esse instrumento poupa a quem faz as buscas o esforço de imaginar todas as palavras capazes de expressar determinada idéia. O investimento na elaboração de um instrumento auxiliar como esse resulta em importante economia num sistema onde haja um grande número de buscas. Esse tipo simplificado de tesauro é uma espécie de vocabulário controlado, em que o controle é feito na saída e não na entrada do sistema. É um vocabulário pós-controlado.

Um exemplo esclarecerá ainda mais sobre as propriedades do vocabulário pós-controlado. Imaginemos uma base de dados sobre negócios públicos indexada com um tesauro que inclui o termo *companhias de aviação*, o que permite fazer uma busca genérica sobre este assunto. Não é possível, porém, restringir uma busca a determinada companhia de aviação, pois os nomes específicos das empresas não fazem parte do tesauro. Assim, seria impossível restringir uma busca a um tema específico como 'situação financeira da Varig'; o melhor que se pode fazer é recuperar tudo sobre a situação financeira de companhias de aviação. A busca genérica costuma ser fácil no caso de vocabulário pré-controlado, mas certas buscas altamente específicas são praticamente impossíveis.

Em comparação, vejamos uma base de dados alternativa sobre negócios públicos que dispensa indexação, mas permite buscas nos títulos e resumos. Nesta, a recuperação de itens sobre a Varig ou a Swissair provavelmente seria fácil. Mais difícil seria uma busca genérica sobre companhias de aviação. Para fazer uma busca exaustiva, seria preciso recorrer a algo mais do que o termo *companhias de aviação*, utilizando certos sinônimos, como *empresas de transporte aéreo* e os nomes de empresas específicas. A estratégia de busca ficaria assim '*companhias de aviação* ou *empresas de transporte aéreo* ou *Varig* ou *Swissair* ou *Lufthansa* ou...' — talvez uma lista muito extensa. O que a pessoa que faz a busca está fazendo é criar parte de um tesauro pós-controlado. Lamentavelmente, nos serviços de informação atuais, essas entradas de tesauro são raramente retidas e armazenadas depois de terem sido criadas e utilizadas. Numa grande rede, há muita duplicação de esforços. *Companhias de aviação* pode aparecer como faceta de muitas buscas realizadas durante um ano, e o trabalho de elaborar estratégias de busca de diferentes graus de completeza será repetido continuamente. Seria muito mais sensato armazenar isso em forma recuperável para uso futuro.

Um verdadeiro vocabulário pós-controlado consiste em tabelas com nomes e números de identificação que podem ser chamados e consultados pelos usuários de bases de dados em linguagem natural que façam parte de alguma rede em linha. Assim, a pessoa que faz a busca recuperaria a entrada 'companhias de aviação', a entrada 'questões financeiras', etc. As tabelas são mostradas em linha e os termos selecionados a partir delas. Alternativamente, a tabela inteira pode ser incorporada numa estratégia de busca mediante seus números de identificação. Essas tabelas não precisam se limitar a palavras, podendo incorporar fragmentos de palavras. Assim, uma tabela de cirurgia teria o seguinte aspecto: 'cirurg..., operaç..., secion..., ...seção, ...otomia, ...ectomia, ...plastia', etc. Também é possível inserir no vocabulário uma estrutura mínima por meio de remissivas de tabelas afins.

Um sistema baseado em vocabulário pós-controlado oferece todas as vantagens da linguagem natural e muitos dos atributos do vocabulário pré-controlado. Um sistema como esse poderá ter um desempenho melhor do que outro baseado num vocabulário pré-controlado. Voltando a um exemplo anterior, seria possível realizar buscas, com facilidade, sobre companhias de aviação específicas, ou utilizar a tabela de 'companhias de aviação' para formar a classe definida por 'companhias de aviação' no tesauro convencional. Uma das vantagens da linguagem natural é ser independente da base de dados. Assim, uma tabela de 'companhias de aviação' seria aplicável igualmente a todas as bases de dados no vernáculo. É possível imaginar um tesauro em linguagem natural aplicável a várias centenas de bases de dados.

Um bom exemplo de vocabulário pós-controlado foi a base de dados TERM implementada pelo Bibliographic Retrieval Services (BRS) e descrita por Knapp (1983). TERM era uma base de dados formada por tabelas que representavam conceitos, incluindo tanto termos controlados quanto termos em texto livre

necessários à realização de buscas numa variedade de bases de dados das ciências sociais e comportamentais. Na figura 105 está um exemplo de uma dessas tabelas.

```
TI  POVERTY AREAS
ER  POVERTY-AREAS+/
ME  POVERTY-AREAS*.
PS  POVERTY-AREAS. CONSIDER ALSO: GHETTOS.
SO  CONSIDER: SLUM. GHETTO. APPALACHIA.
EN  SLUMS.
FT  POVERTY AREAS. SKID ROW. BOWERY. SLUM. INNER CITY. POOR NEIGHBORHOODS.
    MILIEU OF POVERTY. DEPRESSED AREAS. SLUMS. GHETTOS. GHETTO. GHETTOES.
    APPALACHIA. LOW INCOME AREAS. GHETTOIZATION. STREET CORNER DISTRICT. ETHNIC
    NEIGHBORHOOD. BLACK NEIGHBORHOOD. BLACK COMMUNITY. SEGREGATED
    NEIGHBORHOOD. DISADVANTAGED AREA. BLACK SCHOOL DISTRICTS. MINORITY
    NEIGHBORHOOD. REDLINED AREAS. REDLINING.
```

FIGURA 105
Exemplo de entrada da base de dados TERM

O título (TI) da tabela é POVERTY AREAS [áreas de pobreza]. Este termo é utilizado para recuperar itens sobre este tópico no ERIC (ER), nas bases de dados indexadas com o *Medical subject headings* (ME), e na base de dados PsycINFO (PS), na qual um termo afim é GHETTOS [guetos]. No *Sociological Abstracts* (SO), possíveis termos são SLUM [favela], GHETTO e APPALACHIA,* enquanto um termo ERIC (EN) mais específico é SLUMS. Finalmente, apresenta-se uma lista detalhada de termos afins em texto livre (FT), úteis para uma busca sobre este assunto em qualquer base de dados em língua inglesa. Era possível desenvolver uma estratégia na base TERM, a qual seria salva e executada nas bases de dados bibliográficos posteriormente. Esta base de dados, infelizmente, não existe mais. No entanto, seu desenvolvedor publicou uma versão impressa exaustiva das expressões em texto livre (não os termos controlados). Ela pode ser vista como um tesauro destinado a buscas em textos (Knapp, 1993).

Um vocabulário pós-controlado em determinado campo de especialização é elaborado pelo esforço intelectual de sères humanos, exatamente da mesma forma de um tesauro convencional. Essa tarefa pode ser extremamente simplificada mediante o processamento por computador das palavras que ocorram em bases de dados relevantes, de modo a dar origem a vários níveis de 'associação estatística'. Talvez, no entanto, fosse mais sensato recolher e organizar os 'fragmentos de busca' efetivamente introduzidos pelos usuários de alguns sistemas em linha (um candidato a isso seria qualquer lista de termos alimentados numa relação do tipo OU), produzindo assim uma espécie do 'tesauro em crescimento' imaginado por Reisner (1966), porém sendo-lhe imposto posteriormente algum controle

* Região montanhosa pobre dos EUA, que tem como centro o estado da Virgínia Ocidental. (N.T.)

editorial. Mais recentemente, Besser (1997) analisou a importância de termos atribuídos pelos usuários em futuras aplicações de recuperação.

Outra abordagem possível consiste em construir um tesauro automaticamente com base em relações semânticas encontradas em dicionários que existam em formato eletrônico (Fox et al., 1988; Ahlswede et al., 1988). Anderson e Rowley (1992) descrevem um método de construção de 'tesauros do usuário final' a partir de textos completos.

Abordagens atuais

A década de 1960 assistiu ao começo de uma quantidade incrível de projetos de pesquisa sobre a utilização de computadores no tratamento de textos. Havia várias razões para essa explosão de atividades: as instituições de pesquisa (e os pesquisadores) tinham em mãos recursos instalados de computação que eram caros e para eles buscavam utilidade, havia disponibilidade de financiamento generoso das pesquisas, procedente de muitas fontes governamentais, e o processamento de textos era amplamente considerado como uma tarefa bastante simples para computadores vistos como 'poderosos' (normalmente, o que era tido como maior obstáculo era a obtenção de uma quantidade significativa de texto em formato eletrônico).

Embora a tradução mecânica fosse o principal objetivo de grande parte dessas pesquisas, também estavam sendo investigadas várias abordagens para a recuperação da informação. Os projetos mais ambiciosos no campo da recuperação da informação procuravam desenvolver sistemas de 'perguntas e respostas' ou 'recuperação de fatos' — isto é, sistemas capazes de responder diretamente uma consulta do usuário ao invés de recuperar um texto que poderia ou não conter a resposta, ou, mais comumente, uma referência desse texto.

Naturalmente, os problemas resultaram muito maiores do que fora antecipado, particularmente na área da tradução mecânica, e logo o interesse pelo processamento de textos começou a minguar na comunidade de pesquisa, bem como nas agências de financiamento, embora alguns projetos melhores hajam resistido e, com os anos, revelado notável avanço e oferecido resultados promissores.

A amplitude das pesquisas sobre processamento de textos hoje em dia lembra as atividades da década de 1960 (ver Jacobs (1992a) e Pereira e Grosz (1994) onde se encontram boas sínteses dos trabalhos desenvolvidos na década de 1990). Este aumento de interesse e atividade tem origem no fato de que agora se encontram enormes quantidades de texto disponíveis em formato eletrônico, de que a capacidade de processamento é muito maior e custa muito menos, e de que hoje existem necessidades sentidas de aplicações viáveis de processamento de textos nos setores público e privado (por exemplo, disseminação eficiente de informações na Rede e as exigências de multilingüismo compulsório da Comunidade Européia). As pesquisas atuais procuram desenvolver 'sistemas inteligentes baseados em textos'.

Paradoxalmente, a mera quantidade de textos disponíveis para processamento hoje em dia coloca desafios notáveis, mas também oferece soluções potenciais que não estavam disponíveis há 30 anos para os pesquisadores. Por exemplo, léxicos de radicais ou de significados de palavras podem conter muitos milhares de entradas ao invés de umas poucas centenas (Jacobs e Rau, 1994) e é possível utilizar associações (co-ocorrências) de palavras em significativos corpos de textos com a finalidade de reconhecer expressões importantes ou desambiguar palavras, preliminarmente ao processamento lingüístico mais complexo de análise sintática (Wilks et al., 1992; Haas, 1996). A freqüência de palavras pode também ser usada para atribuir texto a várias categorias (Jacobs, 1992b).

Ademais, pode-se empregar a 'filtragem estatística', baseada na co-ocorrência de determinadas palavras ou radicais, para selecionar aquelas frases que pareçam mais prováveis de ser 'relevantes' para determinada exigência e, assim, a melhor candidata para uma análise mais refinada (Wilks et al., 1992).

Charniak (1995) chamou atenção para a possibilidade de obter 90% de exatidão ao atribuir uma 'etiqueta' morfológica [*part-of-speech 'tag'*] a uma palavra simplesmente com base no caso mais provável (que ocorra com maior freqüência) e essa exatidão aumentar em até 95–96% mediante simples verificações de contexto (isto é, procura em palavras adjacentes). Exemplo do método da desambiguação baseada no *corpus* encontra-se em Leacock et al. (1993). Addison (1991) estuda o uso desambiguador do contexto num sistema de recuperação de textos.

Stanfill e Waltz (1992) comparam abordagens atuais mais modernas (que, segundo afirmam, incorporam técnicas de inteligência artificial (IA)) com as de anos anteriores, como se segue:

> A IA da forma como foi formulada no passado está agônica, se é que ainda não morreu; uma nova IA está tomando seu lugar. A antiga IA baseava-se em regras e lógica. A nova IA baseia-se na estatística, porém, não a estatística como era formulada no passado. A prática da própria estatística passa por substancial transformação (p. 215)

e Jacobs (1992a) salienta que as abordagens de hoje em dia extraem "mais força da enorme quantidade de textos armazenados do que de regras artesanais".

As abordagens atuais do processamento de texto podem ser consideradas 'inteligentes' na medida em que os computadores possam vir a 'compreender' o texto.* 'Compreender' significa aqui ser capaz de interpretar o significado de uma frase, sem ambigüidade. Normalmente, isso requer alguma forma de análise sintática. A análise sintática procura identificar o papel de uma palavra numa frase (por exemplo, substantivo ou verbo), reconhecer os diferentes elementos estruturais (oração substantiva, oração verbal, oração prepositiva, e assim por diante), e assim determinar as diversas funções dentro de uma frase (por exemplo, sujeito, predicativo do sujeito, objeto, predicativo do objeto).

---

* Embora a palavra 'inteligente' possa ser também atribuída ao processo, se ele realizar uma tarefa para cuja execução os seres humanos precisariam de inteligência.

O processamento inteligente de textos vem sendo utilizado, experimental ou operacionalmente, em várias aplicações, inclusive categorização de textos, extração de textos, sumarização e ampliação [*augmentation*], geração de textos, e recuperação otimizada da informação [*enhanced information retrieval*], bem como tradução mecânica.*

O propósito de aplicar métodos mais complexos de processamento da linguagem natural [PLN] às buscas em texto completo foi explicado por Strzalkowski et al. (1999) da seguinte forma:

> a principal motivação deste projeto foi demonstrar que um PLN robusto, ainda que relativamente superficial, pode ajudar a extrair uma melhor representação de documentos textuais para fins de indexação e busca do que quaisquer métodos baseados em palavras simples ou seqüências de palavras comumente adotados em recuperação estatística em texto completo. Isso se baseou na premissa de que o processamento lingüístico pode descobrir certos aspectos *semânticos* do conteúdo dos documentos, algo que a mera contagem de palavras não pode fazer, levando assim a uma representação mais precisa (p. 113-114).

Importante abordagem para lidar com a recuperação de textos, utilizada por vários grupos de pesquisas que atuam no âmbito do TREC, é a *extração de sintagmas* [*phrase extraction*] — isto é, reduzir o texto completo a um conjunto de sintagmas que tenham significado. Um dos motivos para isso está simplesmente no fato de que um sintagma pode ser 'significativo' mesmo que as palavras que o compõem não o sejam. Assim '*joint venture*' pode ser significativo porque ocorre de modo relativamente infreqüente numa base de dados, embora as palavras componentes ocorram com demasiada freqüência para que sejam consideradas significativas (Strzalkowski et al., 1999). Foram adotados muitos métodos de extração de sintagmas. Um deles, o método 'núcleo + modificador' ['*head + modifier*'], emprega análise sintática e subseqüente normalização para, por exemplo, reconhecer que '*weapon proliferation*' e '*proliferation of weapons*' [proliferação de armas] são equivalentes (Strzalkowski et al., 1999).

Grande parte dos trabalhos em curso nesta área procura reduzir um texto completo a uma forma mais breve, mediante algum tipo de extração ou sumarização, visando à recuperação da informação. Essas abordagens são tratadas no capítulo seguinte, que também procura avaliar o que elas chegaram a concretizar. Este capítulo limitou-se às buscas em textos de per si, ao invés dos métodos automáticos de indexação ou sumarização, embora essa distinção nem sempre seja fácil de manter, e os capítulos 14 e 15 estão intimamente relacionados.

---

* Em algumas aplicações de processamento de textos é necessário que o computador possa distinguir entre componentes lógicos do documento (por exemplo, título, resumo, texto principal, notas de rodapé, tabelas, figuras) e identificar relações entre eles (como a ordem de leitura). Isso foi denominado, de forma um tanto empolada, 'compreensão do documento' (ver, por exemplo, Semeraro et al., 1994, e *Proceedings of the Third International Conference*, 1995).

As buscas em textos baseiam-se, em geral, em textos em formato eletrônico criados a partir do teclado de um computador ou convertidos do formato impresso por meio de leitoras de caracteres ópticos (embora possam também derivar de entrada falada, como vimos no capítulo 13). Algumas pesquisas foram feitas sobre buscas e recuperação de documentos manuscritos (ver, por exemplo, Perrone et al., 2002), embora não haja clareza sobre quais seriam suas aplicações potenciais.

O que foi concretizado?

Embora as revistas profissionais populares continuem a fazer afirmativas bastante entusiásticas, os autores sérios são muito mais realistas acerca do que já foi conquistado em matéria de processamento automático de textos. Knight (1999), por exemplo, nos diz que:

> As aplicações de linguagem natural, como a tradução mecânica, reconhecimento da fala, recuperação da informação e sumarização, alcançam hoje uma faixa maior de usuários. Quem já usou esses produtos sabe quão imperfeitos eles são. Apesar disso, as pessoas os utilizam porque estão ansiosas em busca de soluções para organizar e pesquisar a enorme quantidade de informações colocadas à sua disposição em linha, em formato textual (p. 58).

Voorhees (1999), que participou dos trabalhos das TRECs durante vários anos, afirmou que as abordagens mais complexas da recuperação da informação a partir de textos produziram resultados desapontadores:

> Atualmente, os métodos de recuperação de uso geral mais bem-sucedidos são os métodos estatísticos que tratam o texto como se não passasse de um saco de palavras [...] as tentativas para melhorar o desempenho da recuperação por meio de processamento lingüístico mais complexo foram em grande parte mal-sucedidos. Na realidade, a menos que seja feito com cuidado, esse processamento pode rebaixar a eficácia da recuperação (p. 32).

No entanto, ela de fato sugere que os níveis mais elaborados de processamento de textos podem ser úteis em atividades de perguntas e respostas e sumarização de documentos.

Strzalkowski et al. (1999) salientam que:

> até o emprego das mais rápidas ferramentas de análise sintática está forçando gravemente os limites da praticabilidade de um sistema de recuperação da informação por causa do aumento da demanda por potência e armazenamento (p. 117-118).

Segundo eles, não passa de modesta a perspectiva de êxito de métodos mais complexos de processamento de texto:

> A principal observação a fazer é que até agora não se comprovou que o processamento de linguagem natural fosse tão eficaz quanto se esperava [...] para conseguir melhor indexação e melhores representações com termos das consultas. O emprego de termos lingüísticos, como expressões, pares de núcleo–modificador, nomes ou mesmo

conceitos simples, ajuda de fato a melhorar a precisão da recuperação, mas os ganhos permanecem muito modestos (p. 143).

Posteriormente, Carballo e Strzalkowski (2000) admitiam que:

As técnicas de processamento de linguagem natural (PLN) podem conter um tremendo potencial para superar as impropriedades dos métodos exclusivamente quantitativos de recuperação de informação textual; no entanto, a prova empírica que sustente essas previsões foi até agora inadequada, e têm demorado a surgir avaliações em escala que sejam apropriadas (p. 155).

Blair (2002) sustenta que as alegações de que houve grande melhoria nos resultados das TRECs ao longo dos anos talvez sejam muito exageradas. Em particular, ele critica os métodos TREC para o cálculo da revocação (uma abordagem que adota uma revocação relativa):

O segundo efeito de estimativas de revocação que não são confiáveis diz respeito ao avanço do campo da Recuperação da Informação como disciplina científica. Isto é, para que avancem as pesquisas sobre recuperação de documentos, temos de conhecer, com total precisão, onde nos encontramos agora. Qualquer incerteza importante na comparação de técnicas de recuperação solapa nossa percepção do que *realmente* funciona e do que não funciona, o que, por sua vez, nos deixa sem qualquer motivo lógico para escolher uma técnica e não outra. Atualmente, a maior parte das técnicas de recuperação automatizada usadas pelos pesquisadores associados às TRECs funciona exatamente no mesmo nível modesto de revocação e precisão. Um dos resultados esperados de estimativas mais exatas de revocação seria o descobrimento de diferenças maiores no desempenho dos sistemas. Então, deveríamos realmente começar a construir sobre os sucessos de algumas técnicas e evitar a perda de mais tempo com outras que são infrutíferas (p. 449).

Saracevic et al. (2003) e Sparck Jones (2003) refutaram algumas críticas de Blair, afirmando (por exemplo) que a avaliação feita sob condições cuidadosamente controladas, baseada em coleções de teste, é essencial para fazer avançar a compreensão dos fenômenos ligados à recuperação; que os resultados desses experimentos podem ser transpostos para serviços de recuperação reais; que não é preciso uma medida de revocação absoluta para comparações controladas do desempenho de diferentes processos de busca; e que, no ambiente controlado das pesquisas TREC, é possível documentar melhorias importantes no desempenho da recuperação à medida que se aperfeiçoam os processos de busca.

Alhures, Sparck Jones afirmou coerentemente que os métodos mais complexos de processamento lingüístico são difíceis de justificar em aplicações voltadas para a recuperação. Depois de passar em revista o estado atual do processamento lingüístico de textos com a finalidade de recuperar informação (ela chama isso de 'indexação lingüisticamente motivada'), conclui (Sparck Jones, 1999) que não está provada sua superioridade em comparação com a abordagem muito mais simples de combinar palavras do texto numa estratégia de busca:

Parece que o efeito de coordenação, otimizado pela redundância da indexação com termos simples, pode bastar para a desambiguação de sentido, pelo menos no caso de bases de dados monolíngües, embora continue em aberto a questão da necessidade de desambiguação explícita em buscas em várias línguas em bases de dados multilíngües. Mesmo quando a discriminação de sentido acrescenta algo ao desempenho [...] isso pode ser obtido mais com métodos estatísticos do que lingüísticos (p. 21).

Ao fazer uma revisão das atividades dos grupos TREC até a TREC-6 (1997), ela (Sparck Jones, 2000) conclui que "métodos baseados na estatística têm desempenho tão bom quanto quaisquer outros, e que a natureza e o tratamento dado ao pedido do usuário são, de longe, o fator dominante no desempenho". Os métodos estatísticos incluem ponderação de termos, expressões simples bem como palavras simples, expansão da consulta e retroalimentação de relevância.

Smeaton (1999) sugere que o processamento lingüístico, embora necessário para aplicações que sejam "exatas e precisas, como a tradução mecânica", constitui ferramenta demasiadamente sutil para a recuperação da informação que ele considera "não uma aplicação exata, e a aproximação é inerente a seu funcionamento devido aos inúmeros graus de incerteza presente nos processos envolvidos".

Além disso, níveis complexos de processamento da linguagem ainda são caros. Em geral, o processamento automático de texto requer a preparação bastante extensa de um programa de computador. Isto é, o programa processa o texto para fazer o que lhe é solicitado, e a saída é vista e corrigida por pessoas, o que leva a alterações do programa. Esse processo iterativo de ensaio e erro continua até o programa obter resultados 'satisfatórios'. Knight (1999) chamou atenção para o volume de processamento exigido para preparar um programa que execute uma tarefa que seres humanos inteligentes executam facilmente. Por exemplo, retirados de um texto os artigos definidos e indefinidos, seria possível escrever um programa capaz de substituí-los. No entanto, Knight afirma que para conseguir um desempenho apenas 'razoavelmente bom' seria preciso o processamento de 20 milhões de palavras de texto em inglês. E acrescenta:

A análise sintática de um texto sem limitações é tarefa excessivamente difícil, devido às ambigüidades em partes da fala (substantivo, verbo, etc.) e da estrutura [...] Mas, apesar de haver algoritmos de aprendizagem promissores, ninguém conseguiu ainda extrair de bases de textos sem tratamento elementos [*parses*] sintáticos que tivessem alguma exatidão (p. 59-61).

Embora o processamento mais complexo da linguagem possa não ser necessário na recuperação de textos, pode sê-lo em aplicações mais exigentes, como a de perguntas e respostas.

Perguntas e respostas

Em setores muito limitados seria possível desenvolver sistemas que realmente respondam perguntas feitas pelo usuário ao invés de simplesmente apontar fontes potenciais onde seriam encontradas as respostas. Sistemas desse tipo seriam particularmente adequados para bases de conhecimento que fossem es-

táticas ou que mudassem muito lentamente. Por exemplo, seria possível desenvolver uma base de dados de óperas, a fim de responder perguntas sobre enredos, cenários, personagens, compositores, estréias, etc. Embora os estudos sobre desenvolvimento de sistemas de perguntas e respostas em campos muito restritos remontem a muito tempo (por exemplo, Green et al., 1963), as tecnologias modernas tornam-nos muito mais viáveis. Por exemplo, Stock (1993) descreve um sistema de hipermídia, o ALFRESCO, com imagens de afrescos italianos do século XIV, capaz de responder ampla variedade de perguntas, inclusive a identificação de personagens ou objetos presentes em certas pinturas. Outro exemplo é encontrado no trabalho de Kupiec (1999). A abordagem ali descrita pode montar 'texto de resposta' a partir de vários documentos diferentes.

Clarke et al. (2001) descrevem processos de resposta automática de perguntas do tipo fatual por meio da Rede. O método envolve a localização e extração de textos que provavelmente contêm a resposta, bem como a seleção da resposta que ocorra com mais freqüência em todos os trechos extraídos.

Uma vertente de perguntas e respostas foi introduzida no grupo TREC em 1999 (TREC-8). Este trabalho foi analisado por Voorhees (2001). No entanto, não é exigido dos participantes do grupo TREC que extraiam respostas do texto, mas que recuperem partes do texto que provavelmente fornecerão a resposta.

Descoberta de conhecimento

Importante campo de pesquisa surgido nos últimos anos refere-se a métodos de extração, das bases de dados, de conhecimentos imprevistos. A terminologia da área é estranhamente confusa e incoerente. Uma denominação perfeitamente razoável e clara é 'descoberta de conhecimento'. Já 'mineração' é amiúde usada como sinônimo de descoberta de conhecimento ou, pelo menos, do elemento central dessa descoberta.* Assim, 'mineração de dados' refere-se ao uso (com o objetivo de descobrir conhecimentos novos) de dados numéricos/estatísticos, 'mineração de textos', ao uso de textos, 'mineração da fala', ao uso da fala gravada, e 'mineração da Rede', ao uso de recursos da Rede. Qualquer que seja a denominação, o processo de descobrir conhecimento envolve basicamente a identificação de padrões significativos nas fontes que estejam sendo utilizadas.

A mineração de dados em geral é revista por Benoit (2000) e a mineração de textos por Trybula (1999). O emprego de bases de dados bibliográficos na descoberta de conhecimento é tratado por Qin e Norton (1999), e Munakata (1999) organizou uma série de artigos sobre descoberta de conhecimento.

Fayyad e Uthurusamy (2002) organizaram um número de periódico dedicado quase totalmente aos métodos de mineração de dados. A mineração de dados é feita para encontrar padrões interessantes nos dados. Exemplificam com a localização de produtos comprados juntos com mais freqüência em supermercados. Embora a mineração possa ser feita para testar uma hipótese, é mais útil desenvolver algoritmos de mineração que essencialmente sugerem as hipóteses.

* Freitas (2002) vê a mineração como um componente da descoberta de conhecimento. Esta última denominação inclui o pré-processamento de dados para facilitar a mineração e o pós-processamento do 'conhecimento descoberto', a fim de validá-lo e refiná-lo.

Nasukawa e Nagano (2001) definem a mineração de texto como o "encontro, no texto, de padrões e regras úteis que indicam tendências e características significativas sobre assuntos específicos". Descrevem um protótipo de sistema para mineração de bases de dados textuais em centros de ajuda comerciais [*help centers*] (centros de suporte a clientes), que, segundo afirmam, pode:

> detectar automaticamente defeitos nos produtos; identificar casos que levaram ao rápido aumento do número de chamadas e as razões por trás disso; e analisar a produtividade do centro de ajuda e mudanças no comportamento dos clientes que envolvam determinado produto, sem ler nenhum dos textos (p. 697).

A mineração de texto é também tratada por Knight (1999).

Embora Etzioni (1996) afirmasse que a Rede não é útil em aplicações de mineração (em sua opinião ela é demasiadamente 'dinâmica e caótica'), outros discordam. Pelo menos dois livros sobre mineração na Rede (Chang et al., 2001; Chakrabarti, 2003) foram publicados. O último é mais teórico do que prático e parece que Chakrabarti está interessado apenas em utilizar a Rede para análise de redes sociais. Não está claro que isso seja 'mineração' no sentido com que este vocábulo é comumente empregado.

Em virtude de o descobrimento de conhecimento implicar a extração de informações, há uma relação próxima entre ele e os processos de extração de textos que serão examinados no capítulo seguinte.

Conclusões

Sistemas que dispensam o controle convencional de vocabulário e a indexação feita por seres humanos podem funcionar, e isso foi comprovado ao longo de um período de mais de 40 anos. Todavia, apresentam, de fato, problemas quando da realização de buscas 'conceituais' genéricas. Embora a linguagem natural apresente vantagens explícitas, é claro que aperfeiçoamentos apropriados (uso limitado da indexação e/ou desenvolvimento de recursos auxiliares de busca) provavelmente melhorarão a eficácia dos sistemas de linguagem natural. Ademais, uma vez que a internet fez crescer, em muitas ordens de magnitude, a quantidade de textos acessíveis para pesquisa, tornou-se cada vez mais necessário implementar sistemas que classificarão os itens recuperados segundo uma ordem de 'relevância provável' ao invés de simplesmente dividir os 'recuperados' pelos 'não-recuperados' (Maron, 1988). Não está claro que níveis complexos de processamento de texto (por exemplo, que envolvam análise sintática) sejam necessários para aplicações de recuperação da informação, ainda que o sejam em serviços verdadeiramente de perguntas e respostas e algumas das aplicações examinadas no capítulo seguinte.

Capítulo 15

Indexação automática, redação automática de resumos e processos afins

Uma imagem muito simplificada do problema da recuperação da informação foi apresentada na figura 1. Agora, na figura 106, tem-se uma versão mais complexa. Em essência, o problema consiste em cotejar necessidades de informação com mensagens. Isso só pode ser feito de modo muito indireto. A maioria das mensagens (aquilo que os autores desejam transmitir) aparece como textos (alguns se apresentam em formato de imagem, de som ou outro formato não-textual), enquanto as necessidades de informação se apresentam como pedidos formulados a um serviço de informação. Este cria representações dos textos, armazena-os numa base de dados e oferece um dispositivo que possibilita que sejam feitas buscas nessas representações. A base de dados pode ser armazenada em papel, microimagem ou formato eletrônico, e o 'dispositivo' que possibilita que se façam as buscas pode ser tão simples quanto o arranjo de entradas num catálogo em fichas ou índice impresso, ou tão complexo quanto um computador e um conjunto de respectivos programas. O serviço de informação também cria representações dos pedidos (enunciados de buscas de algum tipo) e os processa em cotejo com a base de dados, para recuperar as representações de textos que coincidam ou mais se aproximem das representações dos pedidos.

As representações de textos consistirão no próprio texto completo, partes dele ou outra forma de representação construída por meios humanos ou automáticos. As representações de pedidos serão termos, termos apresentados em relações lógicas, enunciados textuais ou 'itens' (por exemplo, um sistema permite ao usuário inserir informações de um item cuja relevância seja conhecida, e, em seguida, procurar outros que de algum modo lhe sejam assemelhados).

Temos à nossa disposição vários recursos de ajuda intelectual que assistem na construção das representações (de textos ou pedidos). O mais óbvio deles é o vocabulário controlado convencional, mas também se usam outros instrumentos auxiliares, como o vocabulário pós-controlado mencionado no capítulo 14.

É evidente que são possíveis muitas variações sobre o tema fundamental da figura 106. Por exemplo, em muitas situações o serviço de informação que cria as representações dos textos (isto é, a formação da base de dados) será diferente dos serviços que realizarão as buscas em tal base. Ademais, quem procura informações poderá não delegar a realização da busca a um especialista em informação, mas, sim, assumi-la pessoalmente. Com o surgimento da internet, a maior

parte das atividades de recuperação da informação envolve buscas em textos de sítios da Rede, e as pessoas que precisam de informações fazem elas próprias as buscas ao invés de delegá-las a outrem.

Este diagrama evidencia os problemas da recuperação da informação. Os textos podem não ser representações perfeitas das mensagens (embora este seja, definitivamente, um problema de comunicação, normalmente não é visto como um problema de recuperação da informação) e, conforme vimos desde os capítulos iniciais, as representações dos textos também podem ser imperfeitas. E, por sinal, os pedidos raramente são representações perfeitas das necessidades de informação e os enunciados de busca podem não ser representações perfeitas dos pedidos. Além disso, o referencial ('esquemas') de um solicitante pode não coincidir com o referencial de um especialista em informação ou, na realidade, o referencial dos autores. Considera-se, então, que o problema da recuperação da informação consiste essencialmente em procurar cotejar aproximações de necessidades de informação com aproximações de mensagens. Não é de admirar que os resultados nem sempre sejam completamente satisfatórios.

Como salientou Bates (1986), o problema da recuperação da informação é mais complexo do que aparenta ser; ela o trata como 'indeterminado' e 'probabilístico'. Parece estar em voga concentrar-se mais no lado da saída da atividade (necessidade de informação—pedido—representação) do que no lado da entrada (mensagem—texto—representação), e a hipótese aí implícita seria que o lado da saída é mais 'complexo'. De fato, Belkin e Belkin et al. (1980, 1982) referem-se ao cotejo do 'estado anômalo de conhecimento' de um consulente com o estado mais 'coerente' de conhecimento dos autores. Como foi apontado ainda no capítulo 2, a função do indexador — prever os tipos de consultas para as quais determinado documento provavelmente será uma resposta útil — não é necessariamente mais simples do que a de quem atua como intermediário: compreender quais os tipos de documentos que satisfazem a um solicitante em dado momento.

Seja como for, a figura 106 é apresentada neste ponto precipuamente para exemplificar o fato de que podem ser usados processos algorítmicos em diversas atividades de recuperação da informação, em substituição ao processamento intelectual por seres humanos. Os computadores podem ser aplicados à indexação automática e à elaboração automática de resumos, bem como a outras operações que envolvam a formação de classes de documentos e de termos, ao desenvolvimento de estratégias de buscas e estabelecimento de redes de associações entre termos. Como o diagrama implica, o computador podem, em certa medida, substituir os seres humanos em praticamente todas as atividades exemplificadas. Atualmente, eles não geram, de modo independente, mensagens ou necessidades de informação, a menos que sejam especificamente programados para esse fim por seres humanos, mas talvez chegue o dia em que também farão isso. Uma vez que a indexação e a redação de resumos constituem a preocupação principal deste livro, neste capítulo dar-se-á mais atenção à aplicação de computadores a essas tarefas.

Capítulo 15

Indexação automática, redação automática de resumos e processos afins

Uma imagem muito simplificada do problema da recuperação da informação foi apresentada na figura 1. Agora, na figura 106, tem-se uma versão mais complexa. Em essência, o problema consiste em cotejar necessidades de informação com mensagens. Isso só pode ser feito de modo muito indireto. A maioria das mensagens (aquilo que os autores desejam transmitir) aparece como textos (alguns se apresentam em formato de imagem, de som ou outro formato não-textual), enquanto as necessidades de informação se apresentam como pedidos formulados a um serviço de informação. Este cria representações dos textos, armazena-os numa base de dados e oferece um dispositivo que possibilita que sejam feitas buscas nessas representações. A base de dados pode ser armazenada em papel, microimagem ou formato eletrônico, e o 'dispositivo' que possibilita que se façam as buscas pode ser tão simples quanto o arranjo de entradas num catálogo em fichas ou índice impresso, ou tão complexo quanto um computador e um conjunto de respectivos programas. O serviço de informação também cria representações dos pedidos (enunciados de buscas de algum tipo) e os processa em cotejo com a base de dados, para recuperar as representações de textos que coincidam ou mais se aproximem das representações dos pedidos.

As representações de textos consistirão no próprio texto completo, partes dele ou outra forma de representação construída por meios humanos ou automáticos. As representações de pedidos serão termos, termos apresentados em relações lógicas, enunciados textuais ou 'itens' (por exemplo, um sistema permite ao usuário inserir informações de um item cuja relevância seja conhecida, e, em seguida, procurar outros que de algum modo lhe sejam assemelhados).

Temos à nossa disposição vários recursos de ajuda intelectual que assistem na construção das representações (de textos ou pedidos). O mais óbvio deles é o vocabulário controlado convencional, mas também se usam outros instrumentos auxiliares, como o vocabulário pós-controlado mencionado no capítulo 14.

É evidente que são possíveis muitas variações sobre o tema fundamental da figura 106. Por exemplo, em muitas situações o serviço de informação que cria as representações dos textos (isto é, a formação da base de dados) será diferente dos serviços que realizarão as buscas em tal base. Ademais, quem procura informações poderá não delegar a realização da busca a um especialista em informação, mas, sim, assumi-la pessoalmente. Com o surgimento da internet, a maior

parte das atividades de recuperação da informação envolve buscas em textos de sítios da Rede, e as pessoas que precisam de informações fazem elas próprias as buscas ao invés de delegá-las a outrem.

Este diagrama evidencia os problemas da recuperação da informação. Os textos podem não ser representações perfeitas das mensagens (embora este seja, definitivamente, um problema de comunicação, normalmente não é visto como um problema de recuperação da informação) e, conforme vimos desde os capítulos iniciais, as representações dos textos também podem ser imperfeitas. E, por sinal, os pedidos raramente são representações perfeitas das necessidades de informação e os enunciados de busca podem não ser representações perfeitas dos pedidos. Além disso, o referencial ('esquemas') de um solicitante pode não coincidir com o referencial de um especialista em informação ou, na realidade, o referencial dos autores. Considera-se, então, que o problema da recuperação da informação consiste essencialmente em procurar cotejar aproximações de necessidades de informação com aproximações de mensagens. Não é de admirar que os resultados nem sempre sejam completamente satisfatórios.

Como salientou Bates (1986), o problema da recuperação da informação é mais complexo do que aparenta ser; ela o trata como 'indeterminado' e 'probabilístico'. Parece estar em voga concentrar-se mais no lado da saída da atividade (necessidade de informação—pedido—representação) do que no lado da entrada (mensagem—texto—representação), e a hipótese aí implícita seria que o lado da saída é mais 'complexo'. De fato, Belkin e Belkin et al. (1980, 1982) referem-se ao cotejo do 'estado anômalo de conhecimento' de um consulente com o estado mais 'coerente' de conhecimento dos autores. Como foi apontado ainda no capítulo 2, a função do indexador — prever os tipos de consultas para as quais determinado documento provavelmente será uma resposta útil — não é necessariamente mais simples do que a de quem atua como intermediário: compreender quais os tipos de documentos que satisfazem a um solicitante em dado momento.

Seja como for, a figura 106 é apresentada neste ponto precipuamente para exemplificar o fato de que podem ser usados processos algorítmicos em diversas atividades de recuperação da informação, em substituição ao processamento intelectual por seres humanos. Os computadores podem ser aplicados à indexação automática e à elaboração automática de resumos, bem como a outras operações que envolvam a formação de classes de documentos e de termos, ao desenvolvimento de estratégias de buscas e estabelecimento de redes de associações entre termos. Como o diagrama implica, o computador podem, em certa medida, substituir os seres humanos em praticamente todas as atividades exemplificadas. Atualmente, eles não geram, de modo independente, mensagens ou necessidades de informação, a menos que sejam especificamente programados para esse fim por seres humanos, mas talvez chegue o dia em que também farão isso. Uma vez que a indexação e a redação de resumos constituem a preocupação principal deste livro, neste capítulo dar-se-á mais atenção à aplicação de computadores a essas tarefas.

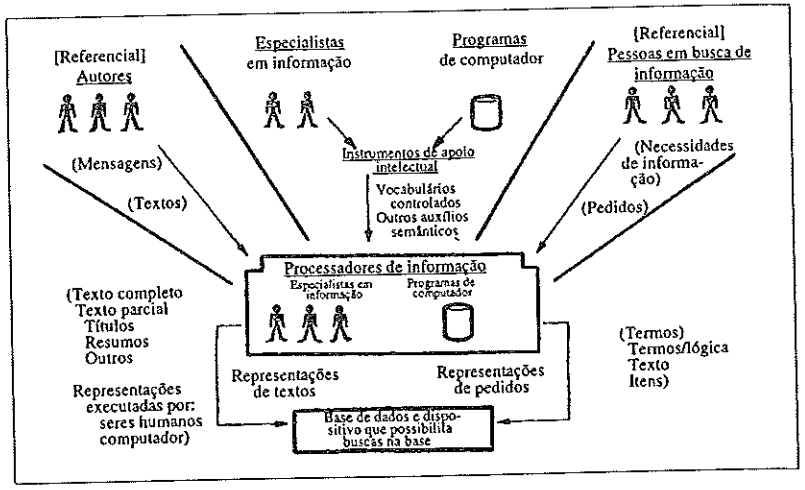


FIGURA 106
Os problemas fundamentais da recuperação da informação

Indexação por extração automática

No começo deste livro fez-se uma distinção entre indexação por atribuição e indexação por extração. A maior parte da indexação feita por seres humanos é por atribuição, pois envolve a representação do conteúdo temático por meio de termos selecionados de algum tipo de vocabulário controlado. Na indexação por extração, palavras ou expressões que aparecem no texto são extraídas e utilizadas para representar o conteúdo do texto como um todo. Os indexadores humanos procurarão selecionar expressões do texto que pareçam ser bons indicadores daquilo de que trata um documento. Provavelmente serão influenciados pela freqüência com que um termo aparece no documento e talvez onde aparece — no título, resumo do autor, legendas das ilustrações, etc. — e por seu contexto.

Admitindo que o texto exista em formato eletrônico, é óbvio que o computador pode ser programado para realizar a indexação por extração, adotando esses mesmos critérios de freqüência, posição e contexto. A indexação automática baseada na freqüência de palavras tem origem na década de 1950 e no trabalho de Luhn (1957) e Baxendale (1958). É possível escrever programas simples para contar as palavras num texto, desde que este tenha sido cotejado com uma lista de palavras proibidas, a fim de eliminar palavras não-significativas (artigos, preposições, conjunções e assemelhados), e, em seguida, ordenar essas palavras segundo a freqüência de sua ocorrência. As palavras do topo da lista serão, evidentemente, escolhidas para serem os 'termos de indexação' do documento. A definição do ponto de corte (ou seja, o ponto em que a lista será interrompida) obedecerá a alguns de vários critérios possíveis: um número absoluto de palavras, um número relacionado com a extensão do texto ou palavras que ocorram com freqüência acima de determinado limiar. Um programa pouco mais complexo extrairá expressões que ocorram com freqüência importante no texto. Assim, um documento poderá ser representado com uma combinação de palavras e expressões, e o critério de freqüência para a seleção das expressões será menos rigoroso do que o critério pelo qual se selecionam as palavras importantes.

Ao invés de selecionar palavras e expressões, os programas podem ser escritos para selecionar radicais. Assim, o radical *calor* seria escolhido e armazenado em vez das variantes *calor, caloria* e *calorimetria*. Empregam-se programas para derivação automática, a fim de eliminar apenas terminações selecionadas de palavras (por exemplo, 'ado', 'ada', 'ando'). Evidentemente, é possível atribuir pesos a todas as palavras, expressões ou radicais, que reflitam a freqüência com que ocorrem no documento. Por exemplo, o radical *calor* pode receber um peso numérico relativo ao fato de aparecer no texto, digamos, 12 vezes.

Os critérios de freqüência podem ser complementados com outros critérios. Por exemplo, Baxendale (1958) propôs que somente a primeira e a última frase de cada parágrafo fossem processadas, pois um de seus estudos demonstrara que a primeira era o 'tópico frasal' em 85% das vezes e a última o era em outros 7% dos casos. Considerava-se 'tópico frasal' aquele que provia o máximo de informações relativas ao conteúdo. Nos primórdios da indexação automática foram propostos ou testados vários outros métodos para identificar os segmentos do texto 'ricos em informação'; programas de computador procurariam certos elementos, como locuções prepositivas, textos que viessem após 'palavras sugestivas', como *conclusões* e *resumo do autor*, e partes do texto que incluíssem as ocorrências primeiras de substantivos.

Uma evidente desvantagem do emprego da freqüência de palavras simples ou expressões para a seleção de termos está em que, mesmo depois de usar uma lista de palavras proibidas, algumas das palavras que ocorrem freqüentemente num documento podem não ser bons discriminantes — que sirvam para diferençar este documento de outros na base de dados — porque também ocorrem com freqüência na base de dados como um todo. Tomando-se um exemplo óbvio, as palavras *biblioteca* e *informação* não seriam muito bons discriminantes de itens numa coleção de biblioteconomia e ciência da informação. Assim, num documento a palavra *biblioteca* ocorre 12 vezes, enquanto a palavra *amianto* só ocorre quatro vezes. No entanto, o último termo é muito melhor discriminante, uma vez que se trata de um termo que raramente ocorre na literatura de biblioteconomia. Seria um termo altamente importante numa coleção deste assunto, mesmo que só ocorresse uma única vez num documento.

A freqüência com que uma palavra ocorre num documento não é a única freqüência para a qual se deve atentar no processamento de textos por computador. A freqüência com que uma palavra ocorre na base de dados como um todo é ainda mais importante. Quer dizer, as palavras que são os melhores discriminantes são aquelas que são imprevisíveis e raras numa coleção — por

exemplo, *amianto* em biblioteconomia, *biblioteca* na base de dados de uma fábrica de cimento-amianto. Na realidade, não é preciso calcular a freqüência com que uma palavra ocorre em toda uma base de dados formada por textos, mas apenas a freqüência com que ela ocorre no arquivo invertido utilizado para executar a busca nos textos (isto é, o número de ocorrências de uma palavra em relação ao número de ocorrências de todas as palavras no arquivo).

Emprega-se, então, ao invés da freqüência absoluta com que uma palavra ocorre num documento, um método de freqüência relativa para a seleção de termos (Oswald et al., 1959). Com este método, selecionam-se palavras ou expressões que ocorram num documento com mais freqüência do que sua taxa de ocorrência na base de dados como um todo. Isso é um pouco mais complicado do que o método de freqüência absoluta, pois exige que se mantenha uma contagem da freqüência com que cada palavra ocorre na base de dados (relativa ao número total de ocorrências de palavras na base de dados), bem como uma comparação dessa taxa de ocorrência com a de uma palavra em determinado documento.

Uma lista de palavras ou expressões extraídas de um documento com base na freqüência relativa será diferente de uma lista criada com base na freqüência absoluta, mas não de forma radical. Muitos dos termos permanecerão os mesmos. Os poucos termos novos serão os que ocorrem raramente no documento, talvez apenas uma vez, mas ainda mais raramente na base de dados como um todo — uma única ocorrência entre as 5 000 palavras de um artigo de periódico é altamente significativa se essa palavra tiver ocorrido até então somente cinco vezes numa base de dados de 10 milhões de palavras! Os termos que desaparecerão, evidentemente, serão os que, embora ocorram freqüentemente num documento, ocorrem freqüentemente na base de dados como um todo.

Evidentemente, os termos selecionados com base na freqüência relativa *não* devem ser radicalmente diferentes dos selecionados com base na freqüência absoluta. Para uma recuperação da informação eficaz precisa-se de termos que sejam bons discriminantes de documentos, e também de termos que formem classes eficazes de documentos. Se for útil mirar exatamente no item raro — o único documento na base de dados que talvez examine os riscos para a saúde do amianto empregado em forros de bibliotecas —, alguém também pode querer recuperar grupos de documentos afins. Palavras como *riscos* ou *perigos* talvez não sejam tão raras numa base de dados de biblioteconomia quanto *amianto*, mas serão úteis para recuperar uma certa classe de documentos que poderão interessar a alguns usuários. Para uma recuperação eficaz da informação, requerem-se, comumente, classes que consistam em mais de um único item.

Os critérios para extrair termos dos documentos incluem, portanto, freqüência absoluta e freqüência relativa, ou uma combinação de ambas, além de critérios posicionais ou sintáticos.* Se se adotar um método relativo para a seleção

* Para um exame completo dos vários critérios adotados para a seleção de termos com base na freqüência de ocorrência, ver Salton e McGill (1983).

de palavras, as listas de palavras proibidas, é claro, não serão necessárias: preposições, conjunções e artigos ocorrerão com freqüência nos itens específicos, mas também em toda a base de dados, e serão assim rejeitadas, junto com palavras significativas mas de ocorrência comum (como *biblioteca* em biblioteconomia).

Os termos também podem ser extraídos do texto quando coincidem com algum tipo de dicionário armazenado de termos 'aceitáveis'. Essa foi a base do importante trabalho sobre indexação com auxílio de computador realizado na década de 1970 pelo Defense Documentation Center (ver, por exemplo, Klingbiel, 1971). Essencialmente, as cadeias de palavras que ocorriam nos títulos e resumos eram cotejadas com uma base de dados em linguagem natural [Natural Language Data Base (NLDB)]. As cadeias de palavras que coincidiam tornavam-se candidatas a termos de indexação. Klingbiel e Rinker (1976) compararam os resultados da indexação com auxílio de computador com os resultados da indexação feita por seres humanos. Como resultado de três estudos de casos, concluíram que a indexação com auxílio de computador e sem revisão posterior alcança níveis de revocação comparáveis aos alcançados pela indexação feita por seres humanos, e que a precisão alcançada pela indexação com auxílio de computador é pelo menos tão boa quanto a alcançada pela indexação feita por seres humanos. A indexação por computador com revisão posterior logrou resultados de revocação comparáveis e melhor precisão do que a indexação feita por seres humanos. Esta abordagem da indexação é atualmente adotada no Center for AeroSpace Information da NASA (Silvester et al., 1993, 1994).

Indexação por atribuição automática

A extração de palavras e/ou expressões dos documentos é tarefa que os computadores executam de modo bastante satisfatório. A extração automática apresenta nítida vantagem em relação à extração feita por seres humanos: é totalmente coerente. No entanto, a maior parte da indexação feita por seres humanos não constitui indexação por extração, mas indexação por atribuição, e a realização desse trabalho por computador é, em geral, mais difícil. A maneira óbvia de executar a indexação por atribuição com o emprego de computador é desenvolver, para cada termo a ser atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer freqüentemente nos documentos aos quais um indexador humano atribuiria esse termo. Esse tipo de perfil, por exemplo, para o termo *chuva ácida* incluiria expressões como chuva ácida, precipitação ácida, poluição atmosférica, dióxido de enxofre, etc.

Se a cada termo de um vocabulário controlado correspondesse um perfil desses, seria possível utilizar programas de computador para cotejar as expressões importantes num documento (essencialmente aquelas que fossem extraídas segundo os critérios de freqüência antes mencionados) com essa coleção de perfis, atribuindo um termo ao documento sempre que o perfil do documento coincidisse com o perfil de termos acima de determinado limiar.

Isso parece relativamente fácil. Na prática, porém, é diferente. Em primeiro lugar, os critérios de coincidência teriam de ser um tanto complexos. Se *chuva ácida* ocorrer dez vezes num artigo de periódico, quase certamente o termo de indexação CHUVA ÁCIDA terá de ser atribuído. Suponhamos, por outro lado, que *chuva ácida* ocorra apenas duas vezes no documento, porém *atmosfera, dióxido de enxofre* e *ácido sulfúrico* ocorram com bastante freqüência. Atribui-se o termo CHUVA ÁCIDA? É evidente que muitas combinações diferentes de palavras ou expressões sinalizam o fato de que determinado termo de indexação será candidato à atribuição. Além do mais, a importância de cada combinação, como preditor de que determinado termo será atribuído, implicaria o emprego de diferentes valores de co-ocorrência. Por exemplo, se as palavras *calor, lago* e *poluição* ocorressem poucas vezes num documento, isso seria o suficiente para levar à atribuição dos termos POLUIÇÃO TÉRMICA e POLUIÇÃO DA ÁGUA. Porém *calor* e *lago*, sem o aparecimento de *poluição*, teriam de ocorrer juntos num documento muitas vezes, antes de POLUIÇÃO TÉRMICA ter assegurada sua atribuição.

A expressão *chuva ácida* apresenta grande probabilidade de ocorrer com freqüência num documento que trate do assunto, de modo que a atribuição correta do termo de indexação CHUVA ÁCIDA talvez não seja tão difícil quanto estariam a sugerir as considerações anteriores. O termo POLUIÇÃO TÉRMICA é mais problemático, pois é menos provável que a maioria dos itens sobre 'poluição térmica' inclua ocorrências freqüentes dessa expressão. Outros termos que um indexador humano atribuiria com grande facilidade quase que resistem à atribuição por computador. O'Connor (1965) analisou alguns problemas concernentes a isso. Um bom exemplo é o termo TOXICIDADE. Um indexador pode, legitimamente, atribuí-lo ao defrontar com esta redação: 'Dois dias depois de a substância haver sido ingerida surgiram diversos sintomas', mas é bastante difícil incorporar num programa de computador todos esses preditores (de que o termo TOXICIDADE deva ser atribuído), mesmo que fossem identificados de antemão.

Devido a esses problemas, as tentativas iniciais de atribuir termos automaticamente não tiveram êxito, mesmo quando estavam envolvidos vocabulários muito pequenos de termos de indexação (por exemplo, Borko e Bernick, 1963). Nos últimos 40 anos, porém, desenvolveram-se processos melhores, e agora é possível executar, com maior chance de êxito, a indexação por atribuição.

A indexação automática e processos afins têm, portanto, uma longa história. No resto do capítulo serão vistos em primeiro lugar outros princípios e abordagens anteriores. Os enfoques mais atuais serão analisadas mais ao final do capítulo.

Estudos anteriores sobre indexação

Van der Meulen e Janssen (1977) relatam uma comparação entre indexação por atribuição automática e indexação manual. Neste caso, comparou-se a indexação humana adotada pelo INSPEC com um esquema de indexação automática que substitui expressões, que ocorrem nos resumos, por 'números conceituais'

extraídos de um 'tesauro' armazenado no computador. Embora os autores digam que a indexação automática deu resultados tão bons quanto os obtidos pela indexação humana, tal conclusão baseou-se nos resultados de apenas duas buscas.

Um dos programas mais complexos de indexação por atribuição automática, desenvolvido no BIOSIS, foi examinado por Vleduts-Stokolov (1987). As palavras que apareciam nos títulos de artigos de periódicos foram cotejadas com um Vocabulário Semântico, formado por cerca de 15 000 termos de biologia, os quais, por sua vez, foram ligados a um vocabulário de 600 Cabeçalhos Conceituais (isto é, cabeçalhos de assuntos relativamente genéricos). Assim, os Cabeçalhos Conceituais podiam ser atribuídos pelo computador com base em palavras/expressões que ocorriam nos títulos. Vleduts-Stokolov relatou que cerca de 61% dos Cabeçalhos Conceituais atribuídos por seres humanos poderiam ser atribuídos pelo computador com base apenas nos títulos. Se se considerassem apenas as atribuições primárias e secundárias (o BIOSIS utilizava um esquema de ponderação de termos de três níveis: primário, secundário e terciário), cerca de 75% das atribuições poderiam ser feitas automaticamente. Na realidade, porém, os programas não alcançaram um nível de desempenho tão elevado. Alcançaram de 80 a 90% de êxito em atribuições primárias e secundárias (isto é, atribuíam de 80 a 90% dos 75% que, teoricamente, seriam atribuídos com base nos títulos), e quase esse nível de êxito em todas as atribuições (ou seja, por volta de 80%, ou um pouco mais, dos 61% de atribuições que ocorreriam com base apenas nos títulos). Em outras palavras, ocorria *subatribuição*; quer dizer, os programas deixavam de atribuir termos que deveriam ser e seriam atribuídos por seres humanos. Ao mesmo tempo, também se verificava *superatribuição*: atribuíam-se termos que não deveriam ser atribuídos. Isso estava na mesma faixa da subatribuição: entre 80 e 90% das atribuições de termos pelo computador eram corretas, no sentido de que indexadores humanos também as teriam feito.

Um método algo similar, descrito por Trubkin (1979), foi adotado para indexar automaticamente os resumos de ABI/INFORM (uma base de dados na área de negócios) no período 1971–77. Construiu-se um 'vocabulário-ponte' com cerca de 19 000 termos que remetiam das expressões dos textos para os termos de um vocabulário controlado. Como bastava uma única ocorrência de um termo num título ou resumo para fazer com que fosse atribuído um termo controlado, os processos de indexação automática tendiam a atribuir mais termos a um item do que o faria a indexação humana (média de 16 por item em contraste com 8–12).

Também similares ao trabalho realizado no BIOSIS são os processos de indexação com auxílio de computador implementados pelo American Petroleum Institute (Brenner et al., 1984). Sua finalidade era desenvolver métodos que permitissem ao computador atribuir os termos controlados do tesauro do API com base nos textos dos resumos. Brenner et al. relatam que uma versão anterior do sistema atribuía somente cerca de 40% dos termos que os indexadores humanos atribuiriam, além de atribuir muitos termos supérfluos. Com os ensinamentos adquiridos nessa experiência, os autores, contudo, sentiam-se otimistas quanto à

possibilidade de os processos informatizados atribuírem cerca de 80% dos termos que deveriam ser atribuídos, e que a isso se seguiria uma redução significativa das atribuições supérfluas. De fato, desde os primeiros testes, ocorreram melhoramentos notáveis. Martinez et al. (1987) analisam esses melhoramentos e também descrevem os problemas encontrados ao fazer a ligação entre expressões dos textos e os termos do tesauro. Posteriormente, Hlava (1992) analisou progressos na abordagem do API no que concerne à ligação de termos de indexação em uma língua com termos de indexação em outra (por exemplo, do inglês com o alemão e vice-versa).

Um método mais elaborado de ligar expressões de textos a descritores foi desenvolvido na Technische Hochschule Darmstadt. Sua descrição mais completa, feita por Knorz (1983), precisa ser complementada com referências posteriores (por exemplo, Fuhr, 1989; Biebricher et al., 1997). O método de Darmstadt, que adota a técnica da ponderação, calcula a probabilidade que um descritor tem de vir a ser atribuído a um item, supondo-se que determinada expressão textual ocorra no título ou no resumo. Como foi dito antes neste capítulo, uma das mais bem-sucedidas aplicações da indexação por atribuição com auxílio de computador encontra-se atualmente em uso no Center for AeroSpace Information (Silvester et al., 1993, 1994), com base no trabalho de Klingbiel.

Apesar de a indexação por atribuição automática ter melhorado consideravelmente nos últimos 40 anos (ver a seção final deste capítulo), ainda não chegamos ao ponto onde termos de um vocabulário extenso (digamos, 10 000 descritores de um tesauro) possam ser atribuídos de modo completamente automático sem intervenção humana. Um estudo feito por Hersh et al. (1993), que trabalhou com textos médicos, afirma ter obtido melhores resultados com buscas em textos simples do que com a ligação de textos aos termos do vocabulário controlado (termos do Unified Medical Language System).

Na realidade, a indexação por atribuição automática se reveste hoje em dia de reduzido interesse, exceto para a produção de índices impressos. Há 30 anos, despertava interesse mais amplo. Como, então, era muito dispendioso armazenar e processar grandes quantidades de texto em computador, justificava-se qualquer método que reduzisse o texto. Hoje em dia, evidentemente, no caso de existir o texto completo de um item em formato eletrônico, ou se existir um resumo adequado, faz pouco sentido pretender indexá-lo, a menos que venha a ser gerada, a partir da base de dados, alguma forma de índice impresso. Não obstante, conforme será visto mais adiante neste capítulo, existem realmente aplicações em que as formas de indexação por atribuição automática ainda são úteis. Ademais, os métodos de indexação por atribuição automática são essencialmente os mesmos usados na categorização (classificação) de textos ou tarefas de encaminhamento de mensagens, a serem examinadas mais adiante.

Uma forma especial de índice impresso é o que aparece no final dos livros. Os trabalhos visando à produção desse tipo de índice por computador também remontam a mais de 40 anos. Artandi (1963) produziu índices de livro por computador no campo da química. Para cada entrada de índice ('termo de expressão') ela criou uma lista de expressões associadas ('termos de detecção'), e a ocorrência de qualquer uma dessas expressões numa página de texto faria com que fosse selecionada uma das entradas de índice para aquela página. Artandi afirmava que um índice assim produzido comparava-se em qualidade a um índice feito por seres humanos, mas custava bem mais caro. Grande parcela do custo correspondia, porém, à transcrição do texto para formato eletrônico. Como hoje praticamente toda impressão de textos é feita a partir de registros eletrônicos, os fatores de custo não mais favoreceriam o esforço intelectual humano. Apesar disso, os problemas inerentes à produção automática de índices de livros são mais difíceis do que sugere o trabalho de Artandi. Mesmo num campo limitado seria preciso um vocabulário muito grande de termos de expressão e, para cada um deles, também seria muito grande o número de termos de detecção possíveis. Ademais, ambos os vocabulários precisariam ser mantidos atualizados para abrigar os novos desenvolvimentos e as mudanças terminológicas nesse campo.

Evidentemente, Artandi procurava fazer a indexação por atribuição. Uma proposta mais fácil seria extrair expressões do texto do livro que fossem adequadas para funcionar como entradas de índice. Earl (1970) descreve um método de elaboração de índices de livros por computador que envolve a extração de sintagmas nominais. Ela afirma que: "Tudo indica ser possível produzir automaticamente índices de livros que sejam satisfatórios, com um trabalho posterior de revisão para eliminar termos supérfluos." Mais tarde, Salton (1989) descreveu como é possível empregar processos de análise sintática para gerar expressões que se prestam ao uso em índices de livros. Por outro lado, Korycinski e Newell (1990) examinam os motivos pelos quais a produção automática de índices de livros é muito mais difícil do que a indexação automática de artigos de periódicos.

A maioria dos sistemas de indexação automática não são realmente 'automáticos', no sentido de que substituem o ser humano pelo computador, mas se destinam a auxiliar o indexador humano. Uma denominação que melhor se ajusta a eles é 'com auxílio de computador'. Em geral, identificam-se dois métodos principais de indexação com auxílio de computador:

1. Utiliza-se o computador para fornecer vários tipos de apresentação e mensagens em linha que ajudam o indexador. Erros cometidos pelo indexador (por exemplo, emprego de termos fora do padrão ou combinações indevidas de cabeçalho principal/subcabeçalho) são reconhecidos em tempo real com imediata notificação ao indexador.
2. Utilizam-se programas de computador para ler o texto (talvez apenas títulos e/ou resumos) e selecionar termos de indexação mediante processos de extração ou atribuição. Os termos assim selecionados são checados por um indexador humano, que acrescenta outros pontos de acesso que os programas não conseguiram atribuir e/ou elimina termos atribuídos erroneamente.

As abordagens atuais são examinadas na seção final deste capítulo.

Outras formas de classificação

Como vimos no capítulo 2, a indexação é uma forma de classificação: a atribuição de um termo a um item coloca-o numa classe junto com outros aos quais o mesmo termo foi atribuído. São possíveis outros tipos de classificação quando há vários dados sobre itens bibliográficos em formato eletrônico. É possível usar processos automáticos para criar classes de documentos ou classes de termos.

Em sistemas 'convencionais' de recuperação, a realização de uma busca é auxiliada pelas associações entre termos estabelecidas pela mente humana, com a ajuda às vezes de relações constantes de um tesauro ou outro vocabulário controlado. Num método mais automático de recuperação — baseado, por exemplo, no cotejo de consultas em linguagem natural com o texto completo de itens, resumos, ou representações de documentos criadas por computador — também convém incorporar processos automáticos para desenvolvimento de relações entre termos, a fim de melhorar a eficácia das buscas. Co-ocorrência é a relação óbvia a ser explorada pelo computador. Quanto mais freqüentemente dois termos ocorrerem juntos (no texto de documentos ou em listas de termos atribuídos aos documentos), mais provável será que tratem de conteúdo temático similar. Levando isso à sua conclusão lógica, se o termo *A* nunca ocorre sem *B* e o termo *B* nunca ocorre sem *A* (o que seria uma situação muito rara), os dois termos são totalmente interdependentes e seriam completamente intercambiáveis nas buscas. Além da associação direta (*X* e *X* tendem a ocorrer juntos), as associações indiretas entre termos podem também ser derivadas com base nos dados de co-ocorrência. Suponhamos que o termo *D* quase nunca ocorra sem *W* numa base de dados e que *T* também tenda a não ocorrer sem *W*, embora *D* e *T* jamais co-ocorram nos documentos. Conclui-se que há uma relação entre *D* e *T*: são relacionados entre si pelo fato de cada um co-ocorrer fortemente com *W*. Muito provavelmente, *D* e *T* são exatamente sinônimos neste contexto: sinônimos costumam não ocorrer um com o outro, ainda que os termos com os quais co-ocorram sejam muito similares. No exemplo hipotético, *D* seria 'delta', *T* 'vôo livre' e *W* 'asa'.

Na realidade, não se calcula o grau de associação entre dois termos com base na freqüência simples de co-ocorrência, mas na de co-ocorrência relativa à freqüência de ocorrência de cada termo. Por exemplo, se os termos *A* e *B* co-ocorrerem 20 vezes na base de dados, enquanto *A* ocorrer 10 000 vezes, e *B* 50 000 vezes, o 'fator de associação' entre *A* e *B* será fraco. Por outro lado, supondo que *A* ocorra 50 vezes, *B* ocorra 25 vezes, e ambos co-ocorram 20 vezes, o fator de associação será grande, pois é muito improvável que *B* ocorra sem *A* e quase a metade das ocorrências de *A* coincida com as ocorrências de *B*. Portanto, a relacionalidade (R) de dois termos é comumente definida pela equação simples

$$R = \frac{a \; e \; b}{a \; ou \; b}$$

Quando R excede algum limiar preestabelecido, os dois termos são aceitos como se fossem relacionados.

Os dados de co-ocorrência são usados de duas formas: 1) desenvolve-se e armazena-se uma rede de associações entre termos, ou 2) identificam-se e armazenam-se classes separadas de termos com base em associações extraídas da rede. No primeiro caso, os termos introduzidos por quem realiza as buscas, em forma de lista ou dentro de um enunciado em forma de expressão ou frase, podem ser processados automaticamente para produzir uma lista expandida de termos de busca. No método desenvolvido por Stiles (Stiles, 1961; Salisbury & Stiles, 1969), os termos acrescentados a uma estratégia de busca são os relacionados de perto com todos os termos da busca original com base na freqüência de co-ocorrência. Por exemplo, *A*, *B* e *C* ocorrem na estratégia original e *X* e *Y* são acrescentados porque tendem a co-ocorrer com todos os três termos iniciais. O processo poderia continuar de modo a introduzir, digamos, o termo *P* porque está associado a *A*, *B*, *C*, *X* e *Y*. Os itens da base de dados podem receber um peso numérico, que reflita o número de termos que coincidem entre item e estratégia de busca e as forças de associação que existem entre esses termos (com base na co-ocorrência), e os itens recuperados podem ser ordenados por peso. É possível, assim, que alguns itens que aparecem no alto da ordenação [*ranking*] não contenham nenhum dos termos com os quais se iniciou a busca.

Na segunda aplicação, qualquer palavra que ocorra num enunciado de busca pode ser substituída pela classe de palavras a que pertence. Isso é automático ou pode ser feito sob controle de quem faz a busca. Os tipos de classes de palavras que podem ser derivadas dos dados de co-ocorrência foram claramente identificados por Salton e McGill (1983). Num deles, chamado *facção*, todas as palavras do grupo são associadas com todas as outras palavras do grupo acima de um limiar escolhido. Num grupo de *ligação única*, por outro lado, cada palavra precisa estar ligada apenas a uma outra palavra do grupo acima do limiar estabelecido.

As classes formadas mediante processos estatísticos serão muito menos puras do que as de um tesauro convencional. Um grupo de palavras que co-ocorram fortemente incluirá relações de gênero/espécie, parte/todo e outras, como no seguinte exemplo:

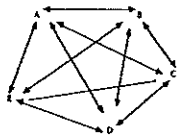
ASA	AERODINÂMICA
AEROFÓLIO	FLUXO
DELTA	
CAUDA	
VIBRAÇÃO	

A pureza da classe não é a questão principal. O que importa é se a classe é potencialmente útil na recuperação. Por exemplo, será provável que a classe hipotética de palavras identificadas acima, se se substituísse automaticamente cada um de seus membros, melhoraria os resultados da busca? Dependendo da busca, parece provável que esse tipo de substituição melhoraria a revocação. Ao mesmo tempo, causaria um grave declínio da precisão, principalmente se a classe (como no exemplo) fosse um conjunto muito heterogêneo de termos.

Salton e McGill (1983) apresentam exemplos de entradas de tesauro extraídas automaticamente de uma coleção de documentos de engenharia (figura 107). Com esse tipo de tesauro, a consulta 'propriedades criogênicas de *x*' seria expandida para '*x* em relação ao conceito 415'. O resultado seriam itens recuperados sobre supercondutividade (isto é, que contêm o radical 'supercondut') de *x*.

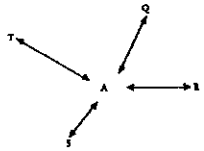
Estas considerações giraram até agora apenas em torno de métodos com os quais se formam classes de termos com base nos documentos onde ocorrem. Os dados que permitem tal classificação são extraídos de uma matriz que mostra quais os termos que ocorrem em quais documentos (matriz termo/documento). É claro que, com esses dados, é também possível fazer a operação inversa. Ou seja, formar classes de documentos com base nos termos que contêm. Salton (1975) e Salton e McGill (1983) identificaram vários tipos dessas classes:

1. A facção



na qual todos os itens *A–E* têm uma forte ligação entre si.

2. A estrela



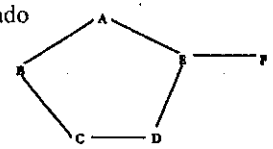
na qual uma classe *AQRST* é definida pelo fato de *Q, R, S* e *T* estarem todos, de alguma forma, ligados de perto a *A*.

3. A fileira

A<__>B <__>C<__> D<__>E

na qual *B* está ligado de perto a *A, C* a *B,* e assim por diante até *E*, o qual não está ligado de perto a qualquer outro item exceto *D*.

4. O conglomerado



que pode ser formado com base em vários critérios. Em geral, no entanto, cada membro se associa aos outros membros do grupo ao alcançar um valor acima de determinado limiar.

Estrelas, fileiras e conglomerados são exemplos dos grupos de ligação única definidos acima.

Uma abordagem muito similar do agrupamento de itens relacionados, chamada 'indexação semântica latente', baseia-se em processo de classificação intimamente relacionado com a análise fatorial (ver, por exemplo, Dumais, 1995).

Também podem ser formadas classes de documentos com base em características não-terminológicas, especialmente várias formas de ligação de citações. As possibilidades disso são exemplificadas na figura 108. Aqui, *X, Y* e *Z* são documentos publicados recentemente que citam os itens anteriores *A, B* e *C*. Uma classe muito simples consistiria em um documento e os posteriores que o citam; por exemplo, *A, X* e *Y*. Como ambos *X* e *Y* citam *A*, existe uma grande possibilidade de que todos os três tenham um conteúdo temático em comum. Isso, evidentemente, é a base da indexação de citações. Ao entrar num índice de citações em *A*, quem realiza a busca encontrará *X* e *Y*, itens estes que citam *A*. Se *A* for um item altamente relevante para os interesses presentes de quem faz a busca, *X* e *Y* também serão relevantes. Se assim for, quem realiza a busca terá logrado êxito sem ter empregado a indexação de assuntos convencional.

Identificam-se outras classes nas relações simples mostradas na figura 108. Por exemplo, considere-se que *X* e *Y* formam uma classe porque ambos citam *A* e *B*. Este é o princípio do *acoplamento bibliográfico* (Kessler, 1962-1965). Quanto mais referências dois (ou mais) itens tiverem em comum, mais forte será seu acoplamento. *X* e *Y* estão fortemente acoplados porque ambos citam *A, B* e *C*. *Z* está menos fortemente acoplado a *X* e *Y* porque tem somente duas referências em comum com estes itens. Outra maneira de dizer isso é que *X* e *Y* formam uma classe forte (de força 3), enquanto *X* e *Z* e *Y* e *Z* são classes fracas (de força 2). É evidente que quanto mais parecidas forem as listas de referências incluídas em duas publicações mais provável será que tratem do mesmo assunto. Assim, se *Q* cita *F, G, H* e *I* apenas, e o artigo *R* também cita somente estes quatro itens, *Q* e *R* quase com certeza tratam do mesmo assunto. Se os dois artigos tiverem essas quatro referências em comum, porém se cada um incluir, digamos, dez referências que o outro não inclui, haverá menos chance de *Q* e *R* tratarem do mesmo assunto, embora a relação entre *Q* e *R* ainda seja considerada muito próxima.

Uma última relação, mostrada na figura 108, é a de *co-citação* (Small, 1973). Afirma-se que os itens *A, B* e *C* formam uma classe porque são citados juntos (co-citados) por *X* e *Y*. Como acontece com o acoplamento bibliográfico, a co-citação pode ocorrer com força variável. Na figura 108, os itens *A, B* e *C* têm uma relação fraca entre si, pois apenas dois itens os citam juntos. Quanto mais itens os co-citarem, supõe-se que mais fortemente relacionados eles estarão.

As classes formadas com base nas ligações de citações apresentam algumas vantagens sobre as classes formadas por meio da indexação de assuntos convencional. O que é mais evidente em tudo isso é que serão independentes de língua e de mudanças terminológicas. O nome de uma doença pode mudar mais de uma vez no decorrer do tempo, porém isto não impedirá que se realize uma busca sobre essa doença num índice de citações, principalmente se o documento inicial que a identifica for do conhecimento de quem faz a busca e se ainda for

408	DISLOCATION	413	CAPACITANCE
	JUNCTION		IMPEDANCE-MATCHING
	MINORITY-CARRIER		IMPEDANCE
	N-P-N		INDUCTANCE
	P-N-P		MUTUAL-IMPEDANCE
	POINT-CONTACT		MUTUAL-INDUCTANCE
	RECOMBINE		MUTUAL
	TRANSITION		NEGATIVE-RESISTANCE
	UNIJUNCTION		POSITIVE-GAP
			REACTANCE
409	BLAST-COOLED		RESIST
	HEAT-FLOW		SELF-IMPEDANCE
	HEAT-TRANSFER		SELF-INDUCTANCE
			SELF
410	ANNEAL		
	STRAIN	414	ANTENNA
			KLYSTRON
411	COERCIVE		PULSES-PER-BEAM
	DEMAGNETIZE		RECEIVER
	FLUX-LEAKAGE		SIGNAL-TO-RECEIVER
	HYSTERESIS		TRANSMITTER
	INDUCT		WAVEGUIDE
	INSENSITIVE		
	MAGNETORESISTANCE	415	CRYOGENIC
	SQUARE-LOOP		CRYOTRON
	THRESHOLD		PERSISTENT-CURRENT
			SUPERCONDUCT
412	LONGITUDINAL		SUPER-CONDUCT
	TRANSVERSE		
		416	RELAY

FIGURA 107
Exemplo de entradas de tesauro extraídas por métodos automáticos
Reprodução de Salton e McGill, *Introduction to modern information retrieval*, 1983, com permissão de McGraw-Hill Publishing Company

citado com freqüência. O princípio do acoplamento bibliográfico pode, naturalmente, ser utilizado para ligar documentos escritos em línguas completamente diferentes; por exemplo, identificar trabalhos em russo e chinês que estejam fortemente acoplados a um trabalho em língua inglesa. Igualmente, uma classe de documentos co-citados incluiria itens em vários idiomas. O que é mais importante, evidentemente, é que as classes formadas por co-citação sofrem mudanças com o passar do tempo, pois novas inter-relações entre os resultados de pesquisas são verificadas por pesquisadores posteriores. Voltando à figura 108, os autores de *X* e *Y* vêem alguma relação entre os itens *A*, *B* e *C*, mas esta relação poderia ter passado despercebida durante muitos anos. *A*, *B* e *C* formam uma classe de itens pela primeira vez em, digamos, 1989, porque foi em 1989 que tanto *X* quanto *Y* foram publicados, porém *A* talvez tivesse sido publicado na década de 1930, *C* na década de 1950 e *B* na década de 1970.

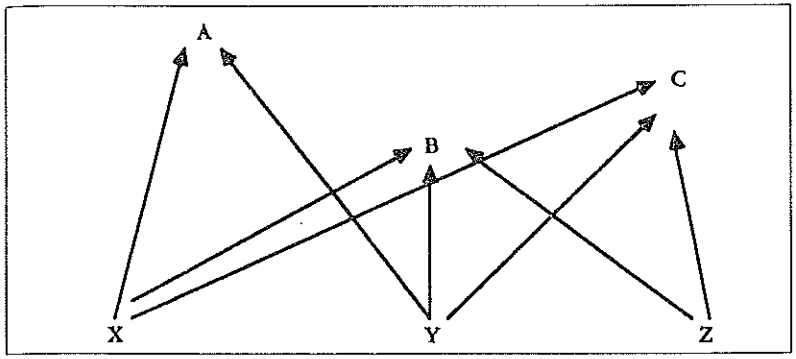


FIGURA 108
Ligações de citações/referências

Os estudos que compararam as classes formadas pela indexação convencional de assuntos com as formadas com base em ligações de citações remontam a cerca de 40 anos (Kessler, 1965). Comparações posteriores incluem Pao (1988), Pao e Worthen (1989) e Shaw (1990b). Uma busca baseada em ligações de citação (citação direta, acoplamento bibliográfico ou co-citação) poderá descobrir itens úteis não encontrados por meio de buscas temáticas convencionais em índices impressos ou bases de dados em linha, porém o método convencional também pode localizar itens que as ligações de citações não conseguem revelar. Os dois métodos são complementares e não concorrentes.

Kwok (1985a,b) menciona o fato de as ligações de referências/citações poderem ser utilizadas na recuperação de informação para formar uma 'coleção ampliada' de itens recuperados. Quer dizer, quando se aplica uma estratégia de busca a uma base de dados da forma normal, empregando palavras do texto ou termos controlados, o conjunto de itens assim recuperados será ampliado com os itens a eles ligados por meio de citações bibliográficas. Ele sugere que o conjunto de termos associados aos itens originalmente recuperados seja ampliado com o acréscimo de termos extraídos dos itens que eles citam. Estes novos termos podem ser termos de indexação atribuídos aos itens citados, ou expressões do texto extraídas dos resumos ou dos títulos. Ele sugere que a ampliação mediante a extração de termos dos títulos dos itens citados é mais praticável. Salton e Zhang (1986) testaram a utilidade de ampliar o conjunto de termos associados aos itens recuperados mediante o acréscimo de palavras do título extraídas de itens 'bibliograficamente relacionados'. As palavras do título foram extraídas de: a) itens citados pelos itens recuperados, b) itens que citavam os itens recuperados, e c) itens co-citados. A conclusão deles é que, embora muitas palavras de conteúdo 'útil' sejam extraídas dessa forma, também serão extraídos muitos termos de utilidade duvidosa, e que o processo não é suficientemente confiável para justificar sua inclusão em sistemas de recuperação operacionais.

É óbvio que as ligações explícitas ou implícitas entre os itens numa rede de hipertexto ou hipermídia são muito similares às ligações de citações aqui examinadas. As implicações para a indexação das ligações de hipertexto/hipermídia são mencionadas no capítulo 16. Um livro organizado por Agosti e Smeaton (1996) é uma boa fonte de pesquisas sobre a utilização de vínculos de hipertexto na recuperação de informações.

Redação automática de resumos

Se os computadores podem ser programados para selecionar termos dos documentos segundo critérios de freqüência, também podem ser programados para selecionar frases dos documentos. Esta é a base do que se denomina comumente 'redação automática de resumos', embora fosse mais exato chamar isso 'elaboração automática de extratos'. Luhn (1958), criador desse método, adotou os seguintes procedimentos:

1. Uma lista de palavras proibidas elimina do processamento ulterior todas as palavras não-significativas.
2. Contam-se as ocorrências de todas as palavras restantes, que são ordenadas segundo sua freqüência de ocorrência (em vez de palavras, podem ser usadas raízes (radicais)).
3. Todas as palavras que ocorram mais de *x* vezes são definidas como palavras de 'alta freqüência' ou 'significativas'.
4. Localizam-se as frases que contenham concentrações dessas palavras de alta freqüência. Consideram-se duas palavras relacionadas dentro de uma frase se não houver mais de quatro palavras intermediárias.
5. Calcula-se um 'fator de significância' para cada frase, da seguinte maneira:
   a) determina-se o número de 'aglomerados' na frase (aglomerado é o grupo de palavras mais extenso, demarcado por palavras significativas, no qual as palavras significativas não se acham separadas por mais de quatro palavras intermediárias);
   b) determina-se o número de palavras significativas no aglomerado e se divide o quadrado desse número pelo número total de palavras dentro do aglomerado;
   c) define-se o fator de significância da frase como o valor do aglomerado mais alto ou como a soma dos valores de todos os aglomerados na frase.

Isso soa mais complicado do que realmente é na prática, e sua explicação fica mais fácil por meio de um exemplo. Vejamos a frase:

A B C D* E F* G* H I J* K L M N O P Q R

onde cada letra representa uma palavra, e as palavras seguidas de asterisco são as consideradas 'significativas'. O aglomerado formado pelas palavras *D–J* contém quatro palavras significativas, de modo que o fator de significância do aglomerado é $4^2/7$ ou 2,3. Este é também o fator de significância da frase, uma vez que ela contém somente um aglomerado.

De acordo com os procedimentos de Luhn, as frases que contenham os fatores de significância mais altos são selecionadas e impressas, na seqüência em que ocorrem no texto, a fim de formar o 'resumo'. É possível estabelecer um ponto de corte, para controlar a quantidade de frases selecionadas. Isso pode basear-se num número fixo de frases ou no número de frases necessárias para atingir certo percentual do texto total do documento. A figura 109 é um exemplo de um 'auto-resumo' produzido de acordo com o método de Luhn.

Ao lidar com documentos muito longos talvez seja conveniente fazer com que os programas selecionem e imprimam frases significativas para cada seção da publicação. Uma vez que os resumos devem salientar a importância específica de um item para a instituição para a qual o resumo é preparado, pode-se incluir uma ponderação adicional numa certa categoria ou lista de palavras, de modo a garantir que as frases que contenham uma ou mais ocorrências dessas palavras sejam selecionadas para inclusão no resumo.

É claro que um resumo montado dessa forma não será muito parecido com um resumo preparado por um ser humano. Uma vez que algumas frases podem vir do primeiro parágrafo, algumas do último, e várias outras talvez do meio do trabalho, o extrato pode parecer bastante desconjuntado. Na realidade, isso não chega a ser de grande importância enquanto as frases escolhidas oferecerem, em conjunto, um quadro exato daquilo de que trata o documento. Alguns pesquisadores, no entanto, discordam disso e insistem para que os extratos obtidos por métodos automáticos apresentem melhor seqüência lógica (Rush et al., 1971, Mathis et al., 1973).

Enquanto Luhn (1959) e Oswald et al. (1959) utilizaram a freqüência de palavras ou expressões para a seleção de frases, outros pesquisadores propuseram ou empregaram critérios alternativos. Edmundson (1969) identificou quatro métodos possíveis:

1. *Método da chave.* Similar ao critério de freqüência de palavras adotado por Luhn. Atribui-se às frases um peso correspondente à soma dos pesos das palavras que as compõem.
2. *Método da deixa.* A presença de certas palavras numa frase sinaliza o fato de que é provável que ela seja um bom indicador de conteúdo. Um 'dicionário de deixas' inclui uma lista de palavras que recebem peso positivo e uma lista de palavras com peso negativo. O valor da significância de uma frase é a soma dos pesos das palavras que a compõem.
3. *Método do título.* A hipótese em que se baseia este método é que as palavras que ocorrem nos títulos e subtítulos são bons indicadores de conteúdo. Atribui-se um valor de significância às frases baseado no número de palavras do título e subtítulo que elas contêm.
4. *Método da localização.* Neste método atribuem-se pesos às frases, tomando por base a posição onde aparecem num documento. As frases que aparecem em certas seções (primeira e última frase dos parágrafos, primeiro e último

Source: *The Scientific American, Vol. 196, No. 2, 86-94, February, 1957*

Title: *Messengers of the Nervous System*

Author: *Amedeo S. Marrazzi*

Editor's Sub-heading: The internal communication of the body is mediated by chemicals as well as by nerve impulses. Study of their interaction has developed important leads to the understanding and therapy of mental illness.

Auto-Abstract*

It seems reasonable to credit the single-celled organisms also with a system of chemical communication by diffusion of stimulating substances through the cell, and these correspond to the chemical messengers (e.g., hormones) that carry stimuli from cell to cell in the more complex organisms. (7.0)†

Finally, in the vertebrate animals there are special glands (e.g., the adrenals) for producing chemical messengers, and the nervous and chemical communication systems are intertwined: for instance, release of adrenalin by the adrenal gland is subject to control both by nerve impulses and by chemicals brought to the gland by the blood. (6.4)

The experiments clearly demonstrated that acetylcholine (and related substances) and adrenalin (and its relatives) exert opposing actions which maintain a balanced regulation of the transmission of nerve impulses. (6.3)

It is reasonable to suppose that the tranquilizing drugs counteract the inhibitory effect of excessive adrenalin or serotonin or some related inhibitor in the human nervous system. (7.3)

*Sentences selected by means of statistical analysis as having a degree of significance of 6 and over.
†Significance factor is given at the end of each sentence.

FIGURA 109
Exemplo de um auto-resumo de Luhn (Luhn 1958)

parágrafo, texto antecedido por entretítulos, como Introdução ou Conclusões) são aceitas como mais indicadoras de conteúdo do que outras.

Descobriu-se que os métodos de deixa, título e localização apresentavam maior probabilidade de concordância quanto às frases a serem selecionadas do que qualquer combinação de métodos que envolvessem o processo de chave, o que levou Edmundson a concluir que este processo, baseado somente nos critérios de freqüência, era inferior aos outros métodos.

Rush et al. (1971) argumentam que qualquer método útil de extração deve incluir critérios tanto para *rejeição* como para *seleção* de frases. O método deles para avaliação de frases leva em conta a 'influência contextual'— uma palavra ou seqüência de palavras, e seu contexto circunjacente, oferecem deixas para a aceitação ou rejeição da frase. O método de extração que descrevem baseia-se no cotejo do texto com uma Lista de Controle de Palavras [Word Control List (WCL)], que inclui uma lista de expressões que, se estivessem presentes numa frase, causariam sua rejeição, e uma lista muito menor de expressões que a levariam a ser selecionada. As expressões de rejeição incluem indicadores de que a frase trata de material relativo a antecedentes e não aos objetivos, métodos e resultados do trabalho. As expressões de seleção são as (do tipo 'este artigo', 'este estudo' ou o 'presente trabalho') que quase sempre significam que a frase trata do tema principal do artigo. São também selecionadas frases que possuam palavras significativas do título do documento. Os critérios de freqüência não são postos de lado, mas usados apenas para modificar os pesos associados às

deixas negativas e positivas no WCL. Os métodos de extração desenvolvidos por Rush et al. ofereciam várias vantagens em relação a processos anteriores, inclusive a capacidade de modificar frases extratadas (por exemplo, pela eliminação de expressões parentéticas).

Outra característica era a 'remissão interfrasal': quando uma frase era selecionada para inclusão num extrato era testada, a fim de determinar se seu significado dependia das frases imediatamente precedentes (por exemplo, por incluir expressões do tipo 'portanto' ou 'por este motivo'). Se o significado fosse assim dependente, as frases precedentes, até um máximo de três, eram incluídas no resumo, mesmo que não atendessem a outros critérios de aceitação. Esse método de extração tem, portanto, o potencial de criar extratos que possuam melhor seqüência lógica do que os obtidos mediante processos menos complexos. Na figura 110 tem-se o exemplo de um extrato produzido segundo os procedimentos de Rush et al. (o sistema de elaboração automática de resumos ADAM).

Mathis et al. (1973) introduziram aperfeiçoamentos nos métodos de extração descritos por Rush et al. Tais aperfeiçoamentos referem-se fundamentalmente às características de modificação frasal e remissão interfrasal dos processos anteriores, e se destinam a produzir representações que sejam mais 'legíveis'.

Earl (1970) realizou experiências a fim de determinar se frases significativas poderiam ou não ser identificadas por meio de análise sintática. A hipótese era que as frases que contivessem certas estruturas sintáticas seriam mais indicativas de conteúdo do que outras. Os resultados não foram promissores, devido principalmente ao grande número de tipos de frases que foram identificados. Um processo mais promissor envolvia o uso de critérios tanto sintáticos quanto estatísticos: identificam-se sintagmas nominais no texto, identificam-se as palavras significativas nos sintagmas, fazem-se contagens de palavras, e as frases são selecionadas com base no número de palavras de alta freqüência que contêm.

Paice (1981) descreveu processos de elaboração automática de extratos baseados na identificação de frases com probabilidade de serem bons indicadores daquilo de que trata um documento (por exemplo, que contivessem expressões como 'o principal objetivo' ou 'descreve-se um método').

Fum et al. (1982) descreveram um método de elaboração automática de resumos no qual, segundo afirmam, processos de análise sintática [*parsing*] e ponderação identificam as informações mais importantes transmitidas num texto, eliminam elementos não-essenciais e reestruturam o restante num resumo condensado e expressivo. Eles apresentam como exemplo a frase

> A necessidade de gerar enorme quantidade adicional de energia elétrica e ao mesmo tempo proteger o meio ambiente é um dos principais problemas sociais e tecnológicos que nossa sociedade terá de resolver em futuro próximo [*sic*]

que se reduz a

> A sociedade deve resolver no futuro o problema da necessidade de gerar energia ao mesmo tempo que protege o meio ambiente.

THE CLAVICHORD AND HOW TO PLAY IT *MARGERY HALFORD, CLAVIER 9(2), 36-41 (1970).* ESSENTIALLY, THE CLAVICHORD IS A SHALLOW RECTANGULAR BOX WHOSE FRAGILE STRINGS, UNDER LIGHT TENSION, ARE STRUNG HORIZONTALLY FROM A SINGLE BRIDGE OVER A THIN SOUNDBOARD. THE KEYS ARE SIMPLE LEVERS WITH A BRASS BLADE CALLED A TANGENT MOUNTED VERTICALLY ON THE FAR END. THE SOUND PRODUCED IS EXTRAORDINARILY RICH IN OVERTONES. THE TONE OF THE CLAVICHORD DOES NOT EXIST READY-MADE AS IT DOES ON THE PIANO AND HARPSICHORD; IT IS FORMED AND SHAPED BY THE FINGER, AS ON A BOWED STRINGED INSTRUMENT, WITH THE RESULT BEING A GENUINE, DIRECT, LIVING "FEEL OF THE STRINGS". AS LONG AS HIS FINGER REMAINS IN CONTACT WITH THE KEY, THE PLAYER RETAINS CONTROL OF THE SOUND. THE CLAVICHORD IS THE LEAST MECHANIZED AND THE MOST RESPONSIVE OF ALL KEYBOARD INSTRUMENTS IN THAT IT MEETS THE PLAYER HALFWAY IN ITS INSTANT AND FAITHFUL TRANSMISSION OF HIS SLIGHTEST MUSICAL INTENTIONS. EMBELLISHMENTS CAN BE PLAYED CRISPLY AND BRILLIANTLY. SHAKES, SNAPS, APPOGGIATURAS, TRILLS, TURNS, MORDENTS, AND SLIDES—ALL SO CHARACTERISTIC OF THE PERIOD WHEN THE CLAVICHORD ENJOYED ITS GREATEST POPULARITY—ARE IDEALLY SUITED TO THE INSTRUMENT'S EXQUISITE CLARITY AND RICHNESS OF TONE. THE ACTION IS SHALLOW AND VIRTUALLY WEIGHTLESS. IT IS A PHENOMENON OF THE DOUBLE—ENDED LEVER THAT THE TONE PRODUCED BY A STRIKING FORCE WILL SOUND BETTER, SWEETER, AND RICHER AT MAXIMUM LEVER LENGTH. FOR THIS REASON, THE KEYS OF THE CLAVICHORD ARE PLAYED AS NEAR TO THE FRONT EDGES AS POSSIBLE. EXCEPT FOR THE PLAYING OF OCTAVES, THE THUMB IS NEVER USED ON A RAISED KEY; DISPLAY PIECES OF A VIRTUOSO CHARACTER ARE GENERALLY UNSUITED TO THE PERSONAL QUALITIES OF THE CLAVICHORD. CRAMER SAYS THAT THE ESPECIALLY REMARKABLE FEATURES OF CLAVICHORD MUSIC ARE FLUIDITY, SUSTAINED MELODY DIFFUSED WITH EVER-VARYING LIGHT AND SHADOW, THE USE OF CERTAIN MUSICAL SHADING AND ALMOST COMPLETE ABSTINENCE FROM PASSAGES WITH ARPEGGIOS, LEAPS, AND BROKEN CHORDS;

FIGURA 110

Exemplo de extrato produzido pelo sistema ADAM de redação automática de resumos
Reproduzido de Mathis (1972) com permissão do Department of Computer and Information Science, Ohio State University

Embora isso seja esplêndido como frase, eles não logram demonstrar que os processos que descrevem produzirão uma condensação expressiva e útil de um artigo inteiro.

Hahn e Reimer (1984) descrevem trabalho voltado para o desenvolvimento de um método, inspirado no conceito de 'sistema especialista', para condensação de textos, em que foi adotada uma base de conhecimento de quadros [*frame knowledge base*] aplicada à análise sintática [*parsing*] de textos. Eles preferem o termo *condensação de textos* a *redação de resumos* porque os métodos podem, em princípio, ser utilizados para criar condensações com vários níveis de extensão e pormenores.

Evidentemente, quanto mais formais e coerentes forem os textos dos documentos, mais bem-sucedidos provavelmente serão os processos de elaboração de extratos. Por exemplo, Borkowski e Martin (1975) alegam ter alcançado mais de 90% de êxito na extração automática de ementas e prescrições exaradas em processos, partindo do texto de decisões judiciais.

As abordagens atuais de extração automática, hoje em dia freqüentemente chamada de 'sumarização de textos', são mencionadas mais adiante neste capítulo. Embora os procedimentos correntes sejam capazes de fazer coisas mais complexas, como a combinação bem-sucedida de frases,* é possível que os critérios relativamente simples introduzidos por Luhn e Baxendale sejam tão bons ou melhores do que quaisquer outros para a seleção prática de frases com probabilidade de serem indicativas do conteúdo do documento. Por exemplo, Hui e Goh (1996) compararam quatro critérios diferentes na preparação de resumos de notícias: método de localização, processo indicativo, freqüência de palavras-chave e palavra-chave do título. O emprego de expressões indicativas (por exemplo, 'em conclusão', 'o objetivo era') para identificar frases significativas deu os piores resultados. O critério simples de freqüência de palavras-chave foi melhor, mas os melhores resultados foram obtidos com métodos que atribuíam peso maior à localização (por exemplo, primeiras frases do parágrafo) ou à seleção de frases que continham maior concentração de palavras que também ocorriam em títulos, entretítulos, legendas ou bibliografias.

Operações 'automáticas' de recuperação

Uma vez que a indexação e a redação de resumos são os temas centrais examinados neste livro, a atenção deste capítulo volta-se para a indexação e a elaboração automáticas de resumos. Todavia, certos métodos automáticos de recuperação da informação guardam com isso uma relação suficiente para justificar que sejam objeto aqui de algumas considerações, ainda que de forma sucinta.

Ao longo dos anos, um dos principais objetivos de inúmeros pesquisadores foi o desenvolvimento de processos que permitiriam que um pedido expresso em texto em linguagem natural fosse cotejado com os textos dos documentos — texto completo, texto parcial ou alguma forma de representação. Considera-se isso como uma espécie de coincidência de padrões: atribui-se aos textos da base de dados um tipo de escore, que reflita o grau com que coincidem com o texto de um pedido, o que permite que sejam apresentados, a quem faz a busca, na forma de uma saída ordenada por provável relevância.

São possíveis vários tipos e níveis de coincidência. Examinemos, por exemplo, o pedido

> Patologia, fisiologia, radiografia e tratamento de pneumonia causada por irradiação ou fibrose pulmonar causada por irradiação

e suponhamos que a base de dados consista em textos de resumos. O método mais simples de pontuar uma coincidência seria aquele que simplesmente levasse em conta quantas palavras do pedido ocorrem num resumo. Assim, um resumo receberia um escore elevado se contivesse as palavras 'patologia', 'fisiologia', 'radiografia', 'irradiação' e 'tratamento' (isto é, cinco das oito ocorrên-

* Johnson et al. (1997) apresentam um bom exemplo de estudo sobre a situação atual da produção de resumos mais inteligíveis por meio de concatenação de frases.

cias de palavras significativas do pedido), embora, evidentemente, seja improvável que possa ser relevante, uma vez que não contém nenhuma das palavras do pedido que são mais discriminantes.

São possíveis muitos refinamentos desse nível rudimentar de estabelecimento de coincidência. Um deles consiste em atribuir a cada palavra um escore que reflita o número de vezes em que ela aparece na base de dados como um todo. Assim, 'fibrose' e 'pneumonia' receberiam escores bastante altos, tendo em vista que provavelmente são menos comuns numa base de dados de medicina do que as outras palavras, mais genéricas, do pedido. Por conseguinte, um resumo que contivesse essas duas palavras receberia um escore elevado, mesmo que não contivesse nenhuma das outras palavras do pedido.

O número de ocorrências de uma palavra num pedido e num resumo também pode ser levado em conta na classificação dos documentos. Segundo este critério, um resumo que contenha diversas vezes a palavra *irradiação* tem a probabilidade de receber um escore elevado porque esta palavra é a única que ocorre mais de uma vez no pedido. No caso de uma base de dados que contenha o texto integral dos documentos, é preciso ter em conta a extensão destes. Do contrário, documentos muito extensos sempre terão uma probabilidade proporcionalmente maior de serem recuperados.

A coincidência pode basear-se em radicais de palavras ao invés de palavras completas. Por este critério, um resumo que inclua as palavras *irradiante* e *irradia*, bem como *irradiação*, obteria um escore elevado em relação ao pedido do exemplo.

Se houvesse no sistema um tesauro criado por computador, seria possível substituir uma ou mais de uma das palavras do pedido pelo grupo existente no tesauro (ver figura 107) e ao qual pertencesse essa palavra. Se ocorresse a substituição das palavras *irradiação* e *pulmonar* do pedido, os pesos dos resumos que contivessem as palavras *pulmões* e *raios* aumentariam notavelmente porque *pulmões* e *pulmonar* pertenceriam ao mesmo grupo do tesauro (junto, talvez, com o radical *pneum*), do mesmo modo que *radiografia, irradiação* e *raios*.

Evidentemente, a coincidência será mais precisa se se basear em expressões e não em palavras simples, pelo que qualquer sistema que coteje o texto de um pedido com os textos dos documentos precisa, definitivamente, ter a possibilidade de realizar buscas em expressões. Os resumos que contenham a expressão 'pneumonia por irradiação' receberão um escore alto em relação ao pedido hipotético, do mesmo modo que aqueles que contenham 'fibrose pulmonar por irradiação'. Os resumos onde houvesse a expressão 'fibrose pulmonar' também receberiam um escore alto, embora com menos probabilidade de ser relevantes, a menos que o aspecto 'irradiação' também estivesse presente. Em posição intermediária entre palavras simples e expressões está o emprego da proximidade de palavras — neste caso a capacidade de atribuir pesos maiores a palavras que apareçam perto uma da outra no texto, embora não necessariamente adjacentes.

Fica evidente, com esta exposição, que podem ser usados diferentes critérios na atribuição de um escore ao texto, a fim de refletir o grau em que ele coincide com o texto de um pedido, e que o escore atribuído pode basear-se em mais de um dos critérios examinados (por exemplo, teria em conta o número de coincidências de palavras ou expressões, bem como o índice de ocorrência dessas palavras ou expressões na base de dados como um todo). Teoricamente, portanto, um sistema 'automático' deve incorporar diversos critérios possíveis para o estabelecimento de coincidência, e permitir ao usuário escolher um deles.

O sistema mais elaborado desse tipo geral é o SMART de Salton, desenvolvido e aperfeiçoado ao longo de um período de mais de 30 anos. Existe uma vasta bibliografia acerca do SMART, e se encontra uma boa síntese em Salton e McGill (1983). Embora os processos hajam sido aprimorados desde que esse livro foi publicado, ainda parece ser a melhor descrição dos princípios básicos. O SMART foi projetado de modo a atribuir pesos numéricos aos itens, a refletir a extensão com que coincidem com os enunciados de pedidos, e a apresentar esses itens ao usuário de acordo com uma ordenação por provável relevância, onde aparecem em primeiro lugar aqueles com pesos maiores. O SMART incorpora diferentes critérios para o estabelecimento de coincidência, inclusive a ponderação de termos, que visa a refletir seu índice de ocorrência numa base de dados, coincidência de expressões, e coincidência baseada em raízes de palavras. Também possibilita a incorporação de um tesauro, o que é obtido mediante uma combinação de processamento por computador e por seres humanos. Outro elemento essencial do SMART é a 'retroalimentação de relevância'. Se, numa saída preliminar, o usuário puder indicar quais os itens que são relevantes e quais os irrelevantes, o sistema recalculará o peso dos itens da base de dados. Consegue-se isso com a redução dos pesos relativos às características dos itens não-relevantes e o aumento dos pesos das características relativas aos itens relevantes. Salton (1989) descreveu como a análise sintática dos textos de capítulos de livros, acompanhada de processos de geração de expressões, pode ser aplicada à produção de índices de final de livros.

Os métodos desenvolvidos por Salton determinam essencialmente a similaridade entre dois textos e expressam essa proximidade como um escore numérico, uma 'medida de similaridade'. Nas operações convencionais de recuperação, mede-se a similaridade entre o texto de uma consulta e textos de documentos numa base de dados, e o escore numérico de similaridade será usado para ordenar a saída. Outras utilizações poderão, porém, ser dadas a essa medida de similaridade dos textos. Por exemplo, é possível medir a proximidade de textos de documentos, o que permitirá a formação de classes de textos similares. Ver, por exemplo, o 'mapa de relações textuais' da figura 111, baseado em Salton et al. (1997). Embora os seis textos representados possam ser considerados semanticamente relacionados, alguns são intimamente relacionados (por exemplo, 17012 e 17016 são fortemente relacionados com um valor de 0,57), enquanto as

ligações entre outros pares são fracas (um valor de 0,09 entre 19199 e 22387 e uma ligação completamente não-significante entre 22387 e 8907). Salton et al. propõem que esses processos de medição de similaridade sejam usados para estabelecer vínculos de hipertexto numa rede de informação. Como será examinado mais adiante neste capítulo, podem também ser utilizados para medir a similaridade entre parágrafos no mesmo texto ('similaridade intradocumental') e isso poderá então ser usado como base para a sumarização do texto.

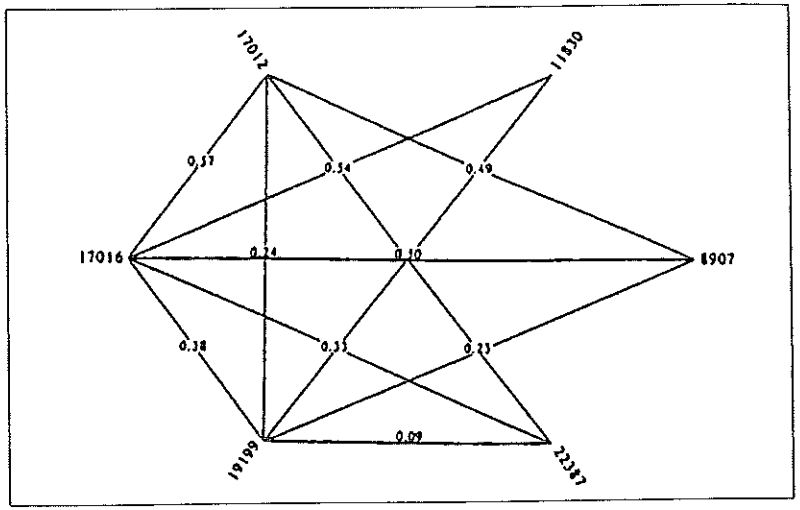


FIGURA 111
Mapa de relações textuais baseado em Salton et al. (1997)
Reproduzido com permissão de Elsevier Science Inc. Os valores numéricos expressam o grau de similaridade entre cada par nos seis textos

Savoy (1995) lida com o estabelecimento de vínculos de hipertexto mediante aplicação de métodos probabilísticos. Também sugere que os vínculos de hipertexto sejam usados para a obtenção automática de novos termos de busca. Por exemplo, se o item *A* for altamente relevante para uma consulta e *A* tiver fortes vínculos de hipertexto com *B*, então *B* poderá também ser relevante. Além disso, os termos fortemente associados com *B* poderão ser úteis para expandir mais a busca.

Outros sistemas também foram desenvolvidos para permitir ao usuário dar entrada a um pedido na forma de enunciado textual. Um exemplo notável foi o sistema CITE desenvolvido por Doszkocs (1983), que também incorpora retroalimentação de relevância. O CITE (Computerized Information Transfer in English [Transferência Computadorizada de Informações em Inglês]) tem sido empregado como interface em linguagem natural com as bases de dados MEDLINE e CATLINE da National Library of Medicine. O CITE funciona numa base de dados

de registros que possuam termos de indexação (como o MEDLINE) ou numa que envolva texto livre (por exemplo, resumos). O sistema pode remover automaticamente os sufixos das palavras (isto é, reduzir as palavras a seus radicais), atribuir automaticamente pesos aos termos da consulta (os pesos refletem a raridade do termo: termos que ocorram raramente na base de dados obtêm peso maior) e apresentar termos possíveis para que o usuário os aprove ou rejeite. Como no SMART, os itens da base de dados recebem um escore numérico que reflete o grau com que coincidem com o enunciado do pedido.

No CITE, os termos relacionados com aqueles empregados na consulta são identificados somente quando a consulta houver sido processada na base de dados. A matéria-prima trabalhada é o conjunto de palavras (termos) relativas aos documentos recuperados. Assim, nos itens recuperados sobre os termos *A*, *B* e *C*, os termos *R* e *T* também podem ocorrer freqüentemente e ser úteis na expansão da busca. Os termos *R* e *T* não são considerados significativos, contudo, a menos que ocorram no conjunto recuperado com maior freqüência do que o esperado. Assim, também se leva em conta a freqüência de ocorrência de um termo na base de dados como um todo. Por exemplo, uma base de dados de biblioteconomia apresenta 85 resumos em resposta a uma consulta simples, como 'avaliação de coleções' (que é interpretada como 'avaliação' *e* 'coleções'). A palavra 'biblioteca' ocorre em 59 desses resumos, mas não é considerada significativa, pois sua taxa de ocorrência no conjunto recuperado (59/85) não excede a taxa de ocorrência na base de dados como um todo. Por outro lado, a palavra 'distribuição' seria considerada associada significativamente com 'coleções' e 'avaliação': ainda que só ocorra em 8 dos 85 resumos, sua taxa de ocorrência (8/85) excede em muito sua taxa de ocorrência na base de dados como um todo.

Uma das grandes vantagens do método de Doszkocs é não exigir um cálculo *a priori* das associações entre termos, uma proposta desanimadora no caso de uma base de dados muito grande. A possibilidade de obter associações úteis entre termos *a posteriori* (depois de a consulta haver sido processada na base de dados), o que requer muito menos processamento do computador, viabiliza processos de otimização das buscas automáticas em sistemas de informação operacionais de porte muito grande. Os sistemas baseados em buscas em linguagem natural e na ordenação por relevância de itens recuperados encontram-se hoje disponíveis comercialmente, como veremos mais adiante.

Método um pouco diferente é adotado no sistema conhecido como Grateful Med (Snow et al, 1986; Bonham & Nelson, 1988). Uma tela formatada em linha convida o usuário a formular sua estratégia de busca. O sistema também sugere ao usuário termos de busca adicionais (extraídos de itens relevantes já recuperados); uma tela de ajuda oferece sugestões para modificação de uma estratégia de busca quando esta não tiver levado à recuperação de qualquer item.

A maioria dos sistemas examinados até agora são sistemas de recuperação bastante convencionais no sentido de que lidam com a busca de registros bibliográficos (ou textos bibliográficos), embora os métodos adotados possam não ser

convencionais. Outros sistemas foram desenvolvidos para buscas de outros tipos de dados. Um exemplo é uma interface em linguagem natural, pouco comum, descrita por Clemencin (1988), que permite a um assinante consultar as 'páginas amarelas' da lista telefônica em linha da França por meio de enunciados de problemas do tipo 'Gostaria de mandar consertar uma câmara fotográfica antiga', 'Preciso contratar um motorista particular', 'Os limpadores de pára-brisas do meu carro estão quebrados', ou 'Torci o tornozelo'. Em resposta, a interface recuperará da lista informações sobre serviços ou profissionais relevantes.

Abordagens atuais

Como foi acima sugerido, a internet provocou tremendo aumento do interesse pelas técnicas de recuperação em geral e pelos métodos automáticos em particular. Alguns sistemas e processos considerados como meramente experimentais há alguns anos são hoje em dia aplicados comercialmente.

Mencionou-se no capítulo anterior que o projeto TIPSTER em muito contribuiu para o progresso alcançado na última década em várias atividades de processamento automático de texto. Este programa, bem como esforços correlatos, incluíram várias conferências sobre recuperação de textos [Text Retrieval Conferences (TRECs)] — a undécima delas realizada em 2002 — bem como conferências sobre compreensão de mensagens [Message Understanding Conferences (MUCs) e, mais recentemente, duas conferências sobre compreensão de documentos [Document Understanding Conferences (DUCs), em 2001 e 2002 (ver http://www-nlpir.nist.gov/projects/duc/). As DUCs tratam da sumarização de textos e são um componente do TIDES (programa Translingual Information Detection, Extraction, and Summarization da DARPA. Também houve uma importante conferência sobre avaliação de métodos de sumarização (Mani et al., 1998).

Embora o patrocínio formal do governo ao TIPSTER haja expirado em outubro de 1998 (Gee, 1999), permanece a cooperação nessas áreas, inclusive com a continuação das atividades TREC. O trabalho do TISPTER e as contribuições das TRECs em especial foram totalmente estudados na literatura (ver, por exemplo, Harman, 1997, Sparck Jones, 1995, e Voorhees e Harman, 1999, 2000). A vertente do TRACK que trata de recuperação interativa foi revista por Over (2001).

Contribuíram também de forma importante para as pesquisas nessa área as conferências sobre processamento de linguagem natural aplicada [Conferences on Applied Natural Language Processing] e as conferências internacionais sobre análise e reconhecimento de documentos [International Conferences on Document Analysis and Recognition].

As atividades de processamento automático de textos relativas ao assunto deste livro incluem indexação com auxílio de computador, indexação completamente automática, encaminhamento de mensagens (categorização de textos), sumarização e extração de textos, e ampliação e geração de textos.

As pesquisas sobre indexação com auxílio de computador, em linha, aplicada a livros, artigos e outras publicações remontam a mais de 30 anos (ver, por exemplo, Bennett, 1969 e Bennett et al., 1972). O auxílio em linha assume várias formas: sugestão de termos aos indexadores (por exemplo, com base no título, resumo ou outro texto trabalhado pelo computador a partir de termos já inseridos pelo indexador), advertência para certos erros do indexador (por exemplo, termos que ainda não se acham no vocabulário do sistema ou combinações indevidas de termos), substituição de termos inaceitáveis por termos aceitáveis, e interface com a base de dados para permitir ao indexador verificar como certos termos foram usados anteriormente ou como certos itens foram antes indexados.

Os sistemas de indexação em linha em ambientes operacionais atuais oferecem vários graus de ajuda e complexidade. Por exemplo, o sistema em uso na National Library of Medicine, o DCMS (Data Creation and Maintenance System), mostra várias mensagens ao indexador, como foi mencionado no capítulo 3.

Sistemas de indexação com auxílio de computador mais complexos superam esses recursos e chegam ao ponto, por exemplo, de indexar parcialmente um item ou, pelo menos, sugerir termos ao indexador. Um deles, o CAIN, foi desenvolvido para ser usado no AGREP, a base de dados da Comunidade Européia sobre projetos de pesquisa agrícola em curso. As descrições dos projetos incluem títulos, resumos e termos não-controlados que indicam o campo de ação do projeto. O CAIN compara esse texto com dois vocabulários controlados (AGROVOC e o CAB Thesaurus) e sugere termos candidatos extraídos dessas fontes (Friis, 1992). Outros sistemas operacionais possuem recursos similares. No caso de sistemas que funcionam com textos curtos (por exemplo, telegramas) e/ou vocabulários controlados relativamente pequenos, sistemas desse tipo são capazes de fazer corretamente grande parte da indexação antes de o indexador humano fazer a revisão para corrigir ou acrescentar o que for preciso.

No Center for AeroSpace Information (CASI) da NASA existe um sistema totalmente operacional, em grande escala, de indexação com auxílio do computador, que foi descrito por Silvester et al. (1994) e Silvester (1998). Uma base de conhecimento constituída de expressões que podem ocorrer na literatura aeroespacial (128 000 entradas em 1998) é empregada para o estabelecimento de ligações com os termos do tesauro da NASA. Isto é, a ocorrência dessas expressões em texto de entrada (normalmente títulos e resumos) leva o sistema a produzir uma lista de descritores candidatos que serão revistos pelo indexador. No CASI, trabalhos relacionados com esse desenvolveram procedimentos para ligação com os termos do tesauro da NASA dos termos atribuídos a registros por outras agências e com o emprego de outros vocabulários (Silvester et al., 1993).

Ainda perdura um grande interesse pela indexação automática destinada a pequenas aplicações especializadas, particularmente no campo biomédico. Em um exemplo (Borst et al., 1992), o texto de resumos de alta de pacientes é analisado, a fim de atribuir automaticamente os descritores clínicos relevantes. De certa forma parecido com esse sistema é o descrito por Oliver e Altman (1994),

que analisará prontuários médicos e a eles atribuirá termos da SNOMED (Systematized Nomenclature of Human and Veterinary Medicine).

Embora se reivindique um nível razoável de desempenho para esse tipo de indexação por atribuição em áreas especializadas, esses processos automáticos geralmente não conseguem alcançar o nível de desempenho obtido por indexadores humanos (ver, por exemplo, Chute e Yang, 1993). Não obstante, esse tipo de indexação automática poderá reduzir a carga de trabalho dos indexadores humanos ao fazer uma atribuição preliminar. Rindflesch e Aronson (1994) analisam alguns dos problemas de ambigüidade presentes na ligação do texto com vocabulários médicos (neste caso, o Unified Medical Language System) e apresentam várias regras de desambiguação.

Está longe de se materializar a indexação por atribuição completamente automática (isto é, sem qualquer intervenção humana) de textos que tenham a extensão de artigos e que tratem de assuntos complexos (por exemplo, em medicina, química ou física), especialmente quando o vocabulário controlado utilizado for muito grande, e por isso foram empreendidas pesquisas para obter sistemas especialistas mais complexos para ajudar o indexador. Um exemplo marcante foi o MedIndEx, que a National Library of Medicine desenvolveu durante muitos anos (Humphrey, 1992). Trata-se de uma abordagem convencional de um sistema especialista baseado em quadros [*frame-based*]. O usuário, que não precisa ser um indexador experiente, mas deve pelo menos ter alguma noção da literatura médica e sua terminologia, é guiado para vários quadros relevantes (por exemplo, tipo de doença, tipo de tratamento) e solicitado a preenchê-los. O sistema pode instar o indexador a atribuir determinado termo e também corrigi-lo quando o termo for empregado de modo inapropriado. Por exemplo, o indexador que atribuir um termo em que apareça a palavra *neoplasia* (câncer) com indicação da localização da doença (por exemplo, *neoplasia óssea*) pode ser lembrado a atribuir um termo associado que representa o tipo histológico da neoplasia (por exemplo, *adenocarcinoma*). Ou o indexador que atribuir uma combinação imprópria, como *fêmur* e *neoplasias ósseas*, poderá ser informado do termo correto, neste caso *neoplasias femorais*. O MedIndEx foi abandonado em favor de pesquisas sobre métodos mais totalmente automáticos.

Outros sistemas especialistas foram desenvolvidos para auxiliar no treinamento de indexadores ao invés de ajudar no processo de indexação de forma rotineira; um sistema desse tipo — CAIT (Computer-Assisted Indexing Tutor) — foi desenvolvido na National Agricultural Library (Irving, 1997).

Qualquer sistema informatizado que auxilie no trabalho de indexação temática pode ser visto como um sistema especialista, pelo menos no sentido mais lato do termo, principalmente se ajudar uma pessoa menos experiente a se aproximar do trabalho de um indexador especializado. E sistemas que sugerem termos aos indexadores ou corrigem certos erros deles podem ser vistos como sistemas que têm pelo menos um tantinho de 'inteligência'.

Alguns sistemas ou programas descritos na literatura são citados como 'artificialmente inteligentes'. Encontram-se exemplos em Driscoll et al. (1991) e Jones e Bell (1992). Os dois últimos autores descrevem um sistema projetado para extrair palavras ou expressões de textos, a fim de formar entradas de índices. Seu funcionamento, em grande parte, baseia-se em listas armazenadas: de palavras a serem ignoradas, palavras/expressões/nomes de reconhecido interesse, e listas auxiliares para desambiguação de homógrafos, para fundir formas do singular/plural e para permitir uma análise simples (lista de terminações de vocábulos). As listas são combinadas para formar um dicionário, que também inclui informações que permitem outros recursos, como, de modo limitado, indexação tanto com os termos específicos quanto com os mais genéricos [*generic posting*].

O sistema descrito por Driscoll et al. também se destina a encontrar no texto termos de indexação úteis. O texto é processado em cotejo com uma lista de mais de 3 000 expressões. A ocorrência de uma delas no texto aciona o uso de regras de inserção e eliminação. As regras de eliminação simplesmente evitam novo processamento de palavras ou expressões que sejam ambíguas, enquanto as regras de inserção podem gerar, por implicação, um conjunto limitado de termos procurados (para completar um 'padrão'). Por exemplo, as palavras '*time*', '*over*' e '*target*' [tempo, sobre, alvo] gerarão AIR WARFARE [guerra aérea], se aparecerem à distância de *x* palavras uma da outra. Malone et al. (1991) apresentam um modelo estatístico para previsão do desempenho deste sistema.

Sistemas como os do tipo descrito por Driscoll et al. e por Jones e Bell são engenhosos. São capazes de realizar indexação por extração, ou extração com atribuição limitada, em nível comparável ao alcançado por indexadores humanos e por um custo menor. No mínimo, são úteis para apresentar termos candidatos que serão revistos por seres humanos. Todavia, não se pode realmente concordar que apresentem inteligência verdadeira. O mesmo se pode dizer dos programas que desenvolvem 'tesauros' e outros recursos auxiliares de busca com base na co-ocorrência de termos (por exemplo, Chen et al. 1995).

Continuam a aparecer na literatura pesquisas destinadas a identificar melhores critérios de associação estatística para a atribuição de termos de vocabulários controlados, com base nas ocorrências de palavras no texto. Plaunt e Norgard (1998), por exemplo, descrevem experiências com a atribuição de termos do tesauro INSPEC com base numa técnica de 'colocação lexical'.

A National Library of Medicine (NLM) investe atualmente expressivos recursos no desenvolvimento de processos para atribuir automaticamente a artigos de periódicos os cabeçalhos do *Medical Subject Headings* (*MeSH*). Isso está se tornando uma necessidade crítica, devido ao volume de processamento: cerca de 400 000 artigos por ano de cerca de 4 300 periódicos biomédicos, com mais de 19 000 termos nos vocabulários *MeSH*. O problema é abordado pela NLM Indexing Initiative. Aron-son et al. (2000) assim a justifica:

À medida que um número cada vez maior de documentos torna-se disponível em

> formato eletrônico e mais organizações desenvolvem 'bibliotecas digitais' para seus acervos, passam a ser necessárias técnicas automatizadas para acessar as informações. Não é possível indexar manualmente cada documento, e novos métodos devem ser desenvolvidos. Essas considerações levaram a promover na biblioteca a Indexing Initiative. Métodos automatizados desenvolvidos e implementados nesse projeto terão um impacto importante na capacidade de a NLM continuar oferecendo serviços de alta qualidade a seu público (p. 17).

Três métodos principais de indexação automática estão sendo pesquisados na NLM. Cada um deles pode gerar uma lista de candidatos a cabeçalhos de assuntos ordenada por relevância provável; alternativamente, a ordenação pode ser obtida pela combinação de dois métodos ou, efetivamente, todos três. Dois desses métodos envolvem a ligação com os termos do *MeSH* de expressões presentes nos títulos dos artigos e nos resumos. O Unified Medical Language System é utilizado como ferramenta para o estabelecimento dessas ligações (ver também Wright et al., 1999, e Aronson, 2001). O terceiro método obtém os termos candidatos mediante o cotejo das palavras, do título e do resumo, de um artigo 'novo' com as palavras que ocorrem no título e no resumo de artigos já indexados. Os termos atribuídos aos artigos coincidentes tornam-se candidatos para atribuição ao novo artigo.

Humphrey (1999) estudou a relação entre as palavras do texto em títulos e resumos de artigos médicos e a categoria de assunto do periódico onde apareciam. Por exemplo, se certo grupo de palavras-chave estiver fortemente associado à categoria 'cardiologia', porque ocorrem freqüentemente em periódicos de cardiologia, o termo CARDIOLOGIA será automaticamente atribuído a qualquer texto onde ocorra esse grupo de palavras-chave. Embora essa categorização genérica não seja adequada para muitas finalidades, poderá ter aplicações práticas. Por exemplo, poderia ser adotada para categorizar automaticamente sítios biomédicos existentes na Rede (Humphrey, 2000; Humphrey et al., 2003).

Outros grupos de pesquisadores, sem afiliação com a National Library of Medicine, desenvolveram métodos de indexação automática por atribuição em biomedicina. Roberts e Souter (2000) descrevem técnicas para atribuição de descritores baseadas em seqüências de palavras de títulos de artigos e ocorrências de palavras em resumos (é preciso que uma palavra-chave ocorra pelo menos três vezes para ser considerada importante). Depois de haver processado 100 registros, a atribuição automática de descritores foi comparada com descritores atribuídos por seres humanos. Os métodos automáticos omitiram muitos descritores que as pessoas atribuíram corretamente e acrescentaram muitos que não deviam ter sido atribuídos, embora também hajam acrescentado uma média levemente superior a um descritor por registro que os seres humanos deveriam ter atribuído mas não o fizeram. Dos 5,5 descritores por registro atribuídos automaticamente, apenas 3,5 foram julgados corretos. As condições em que trabalharam eram muito simples em comparação com as do MEDLINE (por exemplo, um vocabulário muito menor e muito menos termos atribuídos por item) o que serve para dar uma idéia dos grandes problemas envolvidos na tentativa de automatizar totalmente a indexação por atribuição no ambiente de uma base de dados real.

Bradshaw e Hammond (1999) descrevem um sistema em que as citações que uma publicação faz de outra podem levar à extração de texto que seria uma descrição útil para recuperação. Isto é, se a publicação *A* cita a publicação *B*, *A* talvez inclua texto que indica do que trata *B* ou, pelo menos, do que acha que *B* trata. Por exemplo, um trabalho de Harpring (2002) cita um livro de Panofsky e afirma:

> Panofsky identificou três níveis principais de significado na arte: *descrição* pré-iconográfica, *identificação* iconográfica, e *interpretação* iconográfica ou 'iconologia'.

É claro que este texto oferece alguns 'termos de indexação' úteis para Panofsky: significado, arte, iconografia, iconologia e assim por diante. O método é curioso, mas é difícil perceber nele alguma aplicação prática, exceto, talvez, para uma base de dados de textos em área temática altamente especializada. Os exemplos de buscas bem-sucedidas usados por Bradshaw e Hammond (em consultas sobre 'Java' e 'common Lisp') são bastante comuns, principalmente porque resultados iguais teriam sido obtidos com buscas de palavras-chave nos títulos.

Woodruff e Plaunt (1994) descrevem um sistema singular para indexação geográfica automática. Destina-se a:

> [...] extrair de documentos nomes de lugares e também indicadores geográficos mais genéricos, e utilizar a interseção desses referentes para gerar estimativas da área à qual se refere um documento (p. 648).

Nomes de lugares identificados no texto podem ser cotejados com uma base de dados que fornecerá coordenadas de latitude/longitude e também 'características' correlatas, como 'floresta', 'reserva', 'porto' e 'pântano'.

Parece provável que, pelo menos na maior parte das aplicações, sempre haverá itens que não podem ser indexados automaticamente. Ribeiro-Neto et al. (2001), por exemplo, descrevem processos para atribuição automática de categorias da Classificação Internacional de Doenças (CID) a prontuários médicos. O texto dos prontuários é cotejado com termos relativos a cada uma das categorias e subcategorias da CID (extraídas de seu índice, junto com dicionários de sinônimos e siglas). Com base na indexação de mais de 20 000 prontuários, os autores afirmam que obtiveram resultados 'excelentes'. Embora muito poucos dos códigos atribuídos fossem julgados 'errados', mais de 3 000 registros não receberam o código 'ideal'. Desses, 918 não receberam código algum (isto é, o algoritmo não conseguiu indexá-los), que, na grande maioria, asseveram os autores, "representam casos que somente podem ser inteiramente categorizados com auxílio humano (porque, por exemplo, exigem o conhecimento específico de determinada patologia)".

Continuam as pesquisas na área de 'indexação semântica latente'. Anderson e Pérez-Carballo (2001) descrevem o método da seguinte forma:

> A indexação semântica latente (ISL) é um dos mais elaborados esforços atuais visan-

> do a uma indexação automática de alta qualidade. Fundamenta-se em agrupamentos de termos baseados em co-ocorrência e identificação de documentos relativos a tais agrupamentos. Ao se apoiar em dados de co-ocorrência a ISL também consegue lidar com o problema da variedade de termos que expressam idéias semelhantes. [...]
>
> Como exemplo da capacidade de a ISL lidar com terminologia divergente, imaginemos documentos sobre conserto e manutenção de automóveis. Documentos diferentes usarão vários termos diferentes como 'automóvel', 'carro', 'veículo automotor', 'sedã', além dos nomes de marcas e modelos — 'Buick', 'Plymouth', 'Cherokee'. O programa ISL, mui provavelmente, relacionará esses termos entre si devido ao alto nível de co-ocorrência com termos como 'óleo', 'gasolina', 'combustível', 'carburador', 'pneus', 'ar-condicionado', etc. O programa cria agrupamentos de termos altamente relacionados (por meio da co-ocorrência), de modo que, quando um número suficiente deles ocorre num documento, este pode ser ligado ao agrupamento respectivo. Assim, é possível fazer buscas sobre cuidado e manutenção de carburadores de automóveis a gasolina sem nos preocuparmos com as palavras específicas usadas para automóvel. Todas as palavras que significam mais ou menos o mesmo que automóvel serão ligadas ao mesmo agrupamento, à medida que um número suficiente de outros termos co-ocorrentes coincidir com os termos do agrupamento (p. 266).

Na realidade, a indexação semântica latente não é de fato um método de indexação, mas uma maneira de desenvolver automaticamente uma estratégia de busca para produzir termos semanticamente relacionados. Por exemplo, o termo *A* estará um tanto relacionado com o termo *Y* se ambos ocorrerem freqüentemente com o termo *Q*. Com esse método, poder-se-á recuperar documentos possivelmente relevantes cujos termos de indexação diferem dos termos da consulta mas estão estatisticamente relacionados a ele. Segundo Gordon e Dumais (1998):

> Na prática, isso significa que dois documentos que usam vocabulários com alto grau de duplicidade podem ser ambos recuperados mesmo que a consulta somente empregue os termos que indexam um deles. Igualmente, termos serão considerados 'próximos' uns dos outros se ocorrerem em conjuntos de documentos coincidentes (p. 677).

Analisam o emprego desse método como uma maneira de identificar literaturas 'desconexas' (ver, por exemplo, Swanson, 1990): a literatura *A* estará relacionada com a literatura *Y* se os termos de indexação de *A* forem similares aos de *Q* e os de *Y* forem também semelhantes aos de *Q*, embora os termos conectivos em cada caso sejam diferentes. Notem-se as semelhanças entre a indexação semântica latente e a recuperação associativa descrita muito antes por Stiles (1961).

Um importante elemento no processamento automático de texto é o reconhecimento e extração de expressões que provavelmente sejam bons indicadores de conteúdo. As expressões extraídas podem ser empregadas como termos de indexação, ser listadas para formar um tipo de resumo, ou usadas para ligar os termos de um vocabulário controlado. Foram investigados muitos métodos.

Kim e Wilbur (2001) estudaram três diferentes métodos estatísticos para a seleção de expressões portadoras de conteúdo no texto, comparou-as e avaliou seu emprego conjunto na extração de expressões.

Goodby (2001) comparou a extração de expressões por meio de processos lingüísticos (análise sintática para identificar sintagmas nominais com extração baseada em estatística de freqüência, e chegou à conclusão de que o método mais simples de freqüência apresenta resultados tão bons quanto os do método de análise sintática. O método estatístico pode identificar pares de palavras que ocorrem freqüentemente num *corpus*, sua freqüência num documento e sua ocorrência no documento em expressões mais longas (Goodby e Reighart, 2001).

Os processos de indexação automática relacionam-se muito de perto com os processos de categorização de textos (ou melhor, classificação de textos).* Em essência, várias características de um texto, especialmente a ocorrência de diversas palavras ou expressões, são empregadas pelo computador para colocar esse texto numa ou várias categorias preestabelecidas. A origem conceitual disso está nos programas que foram desenvolvidos para a disseminação seletiva de informações (DSI). Nesta, as características de itens publicados recentemente são cotejadas com os 'perfis de interesse' de pessoas ou grupos. Ao ocorrer uma coincidência de determinado valor, o item selecionado será levado ao conhecimento da pessoa ou grupo. Esse tipo de serviço de notificação corrente remonta, de fato, a 1959.

Esse cotejo de documentos recebidos com os perfis de interesses armazenados no sistema é designado 'filtragem e encaminhamento' no ambiente TREC. Robertson (2002) faz uma revisão desse componente das pesquisas TREC.

Uma aplicação importante do encaminhamento é a categorização de notícias recebidas. O sistema CONSTRUE, desenvolvido para a Reuters Ltd., classifica uma seqüência de notícias com o emprego de um esquema de até 674 categorias (Hayes e Weinstein, 1991; Hayes, 1992a). Chen et al. (1994) descrevem processos para identificação de conceitos que ocorrem no texto de reuniões eletrônicas; neste caso, os conceitos são identificados pelos procedimentos ao invés de serem preestabelecidos. Yang (1999) comparou o desempenho de vários métodos de categorização de texto, valendo-se de diferentes critérios de avaliação, em diversas coleções de telegramas de notícias da Reuters.

A categorização automática de texto está incorporada a muitos sistemas operacionais de publicação. Encontra-se um bom exemplo no trabalho de Al-Kofahi et al. (2001). A aplicação inclui a atribuição de ementas de casos jurídicos a um esquema de classificação baseado em mais de 13 000 conceitos legais. A cada semana são produzidas cerca de 12 000 ementas. A categorização baseia-se fundamentalmente nos substantivos e pares de substantivo–substantivo, substantivo–verbo e substantivo–adjetivo que ocorrem no texto da ementa, cotejados com os substantivos/pares de substantivos relativos a cada categoria. A atribuição não é completamente automática — os processos resultam em sugestões de categorização que são examinadas por uma equipe editorial. Afirma-se que os

---

* Ver Guthrie et al. (1999) para uma análise dos critérios de freqüência na categorização de textos.

processos automáticos se comparam favoravelmente com os procedimentos manuais, que substituem, em termos da quantidade de ementas processadas por semana. Para um ingresso semanal de 12 000 ementas a categorização automática faz cerca de 1 600 sugestões, 900 das quais são aceitas, 170 recusadas e 530 não são adotadas por razões editoriais (a precisão é estimada em 89% — 1430/1600).

Há atualmente programas de computador que realizam algum nível de classificação automática de recursos da Rede (Trippe, 2001; Reamy, 2002). Reamy, que trata o processo como 'autocategorização', resume algumas das abordagens:

> A primeira e melhor coisa que um programa de autocategorização pode fazer é examinar com muita rapidez cada palavra do documento e analisar as freqüências de padrões de palavras e, com base numa comparação com a taxonomia existente, atribuir o documento a determinada categoria dessa taxonomia.
>
> Outras coisas que estão sendo feitas com esse programa são 'agrupamento' ou 'construção de taxonomia' em que o programa é simplesmente apontado para uma coleção de documentos, por exemplo de 10 000 a 100 000, e ele pesquisa em todas as combinações de palavras em busca de aglomerados ou agrupamentos de documentos que pareçam ser da mesma classe (p. 18).

Trippe menciona diversos produtos desse tipo, inclusive um da empresa Eprise que é assim descrito:

> De acordo com Hank Barnes, vice-presidente de estratégia da Eprise, 'Um aspecto importante para tornar os conteúdos mais eficazes são as metaetiquetas de classificação. Elas permitem aos usuários de conteúdos encontrar mais facilmente informações relevantes e obter informações mais profundas sobre assuntos específicos'. Barnes observa que a Eprise utiliza esses tipos de etiquetas para localizar informações de modo dinâmico em resposta a ações dos usuários, como seguir determinado caminho num sítio da Rede. Acrescenta Barnes, 'Com freqüência, esse método de fornecimento de conteúdos que se baseia em classificação é muito mais eficaz do que buscas em texto completo ou de utilidade geral' (p. 46).

Kwon e Lee (2003) também tratam da classificação de sítios da Rede, enquanto Lawrence et al. (1999) descrevem procedimentos para citação automática de literatura científica na Rede.

Os processos de categorização de textos até agora descritos representam formas de classificação automática, isto é, a atribuição de itens a classes ou categorias preestabelecidas. Ao longo dos anos, foram feitos estudos sobre a automação do tipo de classificação com o qual os bibliotecários estão mais familiarizados, a saber, a atribuição de números de classificação a livros, mas disso não resultaram sistemas totalmente operacionais. Iyer e Giguere (1995) fizeram estudo sobre o desenvolvimento de um sistema especialista que estabelecesse ligação entre um sistema de classificação e outro, no caso específico do esquema de matemática da American Mathematical Society para a classe de matemática da Classificação Decimal de Dewey. Afirmam que "Uma interface que permita aos matemáticos ter acesso aos acervos de bibliotecas organizados pela Classificação Decimal de Dewey valendo-se do esquema da AMS como interface será certamente útil". Esse tipo de aplicação, porém, parece de utilidade muito limitada.

De interesse mais amplo seria um sistema interativo que ajudasse na atribuição real de números de classificação. Alguns trabalhos nessa linha já foram realizados, mas não em escala muito grande. Por exemplo, Gowtham e Kamat (1995) desenvolveram um protótipo de sistema de classificação no campo da metalurgia com o emprego da Classificação Decimal Universal (CDU). Embora muito menos ambicioso e complexo do que o sistema MEDINDEX antes descrito, o protótipo que descrevem funciona de maneira semelhante, pois sugere ao usuário construir um número de classificação que contenha todas as facetas necessárias (tipo de metal, propriedade, tipo de processo adotado, e assim por diante). Cosgrove e Weimann (1992) também examinam uma abordagem de sistema especialista na utilização da classificação pela CDU, porém de uma perspectiva teórica. Não existe qualquer indício de que algum sistema, mesmo em caráter experimental, haja sido implementado.

Importantes trabalhos sobre classificação automática foram realizados no OCLC. O projeto Scorpion, no OCLC, efetuou experiências com a classificação automática de páginas da Rede com o emprego da Classificação Decimal de Dewey (Thompson et al., 1997). A atribuição baseava-se no cotejo de texto da Rede com as definições textuais dos números de classificação da CDD, mediante o uso de algoritmos desenvolvidos para utilização no sistema SMART de Salton.

Antes, Larson (1992) testou, em pequena escala, a atribuição automática de números de classificação da Library of Congress. Seu objetivo era diferente: a atribuição automática de um único número a um livro com base nos títulos e cabeçalhos de assuntos presentes nos registros MARC. Assim como no estudo feito pelo OCLC, seu algoritmo ordenava os números de classificação em ordem de probabilidade de 'correção'. Larson concluiu que talvez não fosse possível uma classificação totalmente automática, mas uma classificação semi-automática. Isto é, o programa produziria uma lista de números candidatos (os de mais alta pontuação) da qual o classificador selecionaria o que fosse mais apropriado.

Pesquisas sobre classificação automática também são feitas em campos completamente diversos. Por exemplo, Bailin et al. (1993) examinaram trabalhos sobre classificação de componentes de programas de computador (para um repositório de programas reutilizáveis); afirmam que houve características de aprendizado de máquina. Savić (1995) lida com as possibilidades de classificação automática de correspondência administrativa.

Em vários centros de pesquisa, fora do campo da biblioteconomia/ciência da informação, têm prosseguimento trabalhos sobre a construção automática de tesauros. As ferramentas assim construídas, embora, de fato, possivelmente revelem relações úteis entre termos, são muito menos estruturadas do que os tesauros criados por seres humanos. Encontram-se exemplos em Gao et al. (1995), Chen et al. (1995) e Lu et al. (1995).

Embora a indexação assistida por computador possua uma longa história, a redação de resumos assistida por computador (ao contrário dos métodos totalmente automáticos) tem recebido muito pouca atenção. Craven (2000, 2001), no entanto, descreveu um sistema que gerará automaticamente palavras-chave ou expressões a partir de texto completo e as exibirá em janelas para ajudar quem estiver preparando um resumo para esse texto. As expressões são escolhidas com base num escore numérico que reflete o número de palavras-chave 'freqüentes' na expressão, o tamanho da expressão e o número de vezes em que ela ocorre. Os sujeitos de sua experiência julgaram que as expressões extraídas não eram mais úteis do que as palavras-chave na redação dos resumos.

A denominação 'redação automática de resumos' cedeu lugar à denominação 'sumarização de textos'. Na realidade, nenhum grupo de pesquisa conseguiu produzir automaticamente o tipo de resumo que uma pessoa consegue redigir. A sumarização automática ainda é uma questão de seleção de frases e o objetivo das pesquisas nesta área consiste em otimizar essa seleção (no sentido de escolher as frases que melhor representem o conteúdo do texto presente) e organizar as frases selecionadas (possivelmente modificando-as mediante alguma forma de fusão) para melhorar a clareza e utilidade do extrato.

A sumarização pode envolver várias transformações do texto para condensá-lo ainda mais. Por exemplo, é possível agregar enunciados por meio de análise sintática e semântica. Mani (2001) apresenta o exemplo muito simples de 'João e Maria jantaram juntos' e 'Então João lhe propôs casamento' que se agregam para formar 'João propôs casamento a Maria depois do jantar'.

As limitações dos métodos atuais de sumarização foram bem explicadas por Hahn e Mani (2000):

> [...] sua aplicação limita-se à *extração* — selecionar passagens originais do documento-fonte e concatená-las de modo que produzam um texto menor. A *redação de resumos*, em compensação, parafraseia em termos mais gerais aquilo de que trata o texto.
>
> O método de concatenação para fazer a extração em pouco contribui para garantir a coerência do resumo, o que pode dificultar a leitura do texto. Além do que, nem sempre a fonte possui texto — por exemplo, um evento esportivo em videoteipe ou tabelas que mostram dados econômicos — e as ferramentas atuais não podem resumir mídia não-textual. Finalmente, essas ferramentas atualmente não trabalham com fontes múltiplas. Por exemplo, se houvesse muitas notícias na Rede sobre um evento, seria útil se o resumidor pudesse capturar informações comuns e novas (p. 29).

As duas últimas limitações mencionadas realmente não são mais válidas porque agora existem diversos métodos para resumir material em vídeo (ver capítulo 13) e multidocumentos.

Hahn e Mani (2000) salientam que os métodos atuais de extração utilizam um modelo de ponderação linear com vários componentes, tais como localização no texto, número de ocorrências na base de dados como um todo e o aparecimento de expressões-deixa [*cue phrases*]. Assim, uma unidade de texto (geralmente uma frase) seria selecionada com base num modelo do seguinte tipo:

Peso da frase = peso da localização + peso da expressão-deixa + peso da ocorrência no texto + peso da ocorrência na base de dados

Naturalmente, o último componente é um peso negativo: palavras ou expressões obtêm escores mais elevados quanto menos freqüentemente ocorrerem alhures na base de dados. Hahn e Mani também sugerem o emprego de um peso adicional para palavras/expressões baseado na ocorrência alhures no texto (por exemplo, peso maior se também ocorrer no título) ou mesmo ocorrência numa lista de termos que representam interesses atuais.

Salton et al. (1997) descreveram um método de produção automática de resumos de textos completos. Os métodos empregados para medir a semelhança entre pares de documentos (ver figura 11.1) podem ser também empregados para medir a semelhança entre pares de parágrafos no mesmo documento. Assim, podem ser formados agrupamentos de textos, em que um agrupamento consiste em parágrafos, possivelmente extraídos de partes completamente diferentes do texto, que parecem tratar do mesmo tema. Afirmam que isso permite a formação de resumos de textos, inteligíveis, por meio da extração de parágrafos. Observe-se que o trabalho deles é um tanto diferente da maioria dos trabalhos que tratam de redação automática de resumos, que se baseiam na frase como unidade, e não no parágrafo. Os procedimentos empregados por Salton et al. produzem resumos de textos mais longos do que as abordagens mais convencionais.

Resumos de textos produzidos por essa extração de parágrafos foram comparados com resumos produzidos pela extração feita por seres humanos de parágrafos 'importantes'. Os pesquisadores consideram aceitáveis os processos automáticos porque o resumo daí resultante tem tanta probabilidade de coincidir com um resumo extratado por uma pessoa quanto dois resumos extratados por pessoas têm de estar de acordo um com o outro, embora os processos automáticos saiam muito mais baratos.

McKeown et al. (1995) e Maybury (1995) descrevem atividades altamente especializadas de sumarização. Os primeiros geram resumos narrativos de dados armazenados (e não de texto narrativo) relativos a jogos de basquetebol e atividade de planejamento de redes telefônicas, enquanto o sistema de Maybury gera resumos textuais de mensagens militares altamente condensadas e estruturadas (dados de batalha).

Nomoto e Matsumoto (2001) descrevem um método de criação de resumos em que a 'diversidade' é levada em conta na formação do extrato. Isto é, são identificados os vários tópicos abrangidos pelo texto e é selecionada a frase mais representativa para cada tópico.

Saggion e Lapalme (2000) descrevem um método de sumarização baseado em 'análise seletiva'. O método possui duas etapas. Na primeira, um resumo indicativo é apresentado ao usuário (na realidade, apenas uma lista de termos-chave extraídos); se o usuário quiser mais, serão recuperadas e apresentadas a ele passagens importantes do texto.

O método usado por Lehmam (1999) baseia-se na seleção de frases que contenham a maior concentração de palavras 'indicadoras de conteúdo' ou expressões como 'nesta pesquisa', 'o método' e 'é examinado'.

Ainda existe um forte interesse pela preparação automática de extratos. Por exemplo, Moens e Dumortier (2000) descrevem procedimentos para produção de extratos de artigos de revistas de interesse geral. A finalidade desses 'resumos em realce' [*high-light abstracts*] é despertar suficiente interesse dos leitores que, navegando em linha nos resumos, sentiriam vontade de ler o artigo inteiro. Eles descrevem as características almejadas da seguinte forma:

> O resumo em realce é indicativo do conteúdo do texto original. Sugere os principais tópicos do artigo sem entrar em muitos detalhes, o que tornaria supérflua a leitura do texto completo. O resumo em realce possui uma dimensão adicional. Deve não apenas ser factual e sugerir de que trata o artigo, mas também estimular a aquisição do artigo completo. O resumo consiste em recortes de texto, isto é, frases e enunciados extraídos do texto. De preferência contém frases curtas e facilmente inteligíveis, que não dependam do contexto do artigo circundante para permitir uma interpretação correta. É importante incluir linguagem conversacional no resumo (por exemplo, frases em discurso direto, perguntas), porque isso o torna interessante (p. 521).

Seu processo de sumarização utiliza os padrões de discurso característicos dos relatos de notícias, a fim de desenvolver uma 'gramática do texto' que é empregada na análise sintática do texto. A 'sinalização de deixas lingüísticas' identifica frases relevantes para inclusão no resumo.

Moens et al. (1999) sustenta que o conhecimento da estrutura do discurso de vários tipos de textos é útil no projeto de sistemas para geração de texto ou extração de texto. Esses autores trabalham especialmente com textos jurídicos.

Hoje em dia, são comuns os programas para extração automática de frases-chave de textos. Vários produtos encontram-se disponíveis para preparação desses extratos para textos acessíveis na Rede. Diversos produtos, gratuitos ou de baixo preço, foram examinados por Jacsó (2002).

Allan et al. (2001) descrevem métodos para produção de 'resumos temporais' de notícias. A situação é a de uma corrente de notícias sobre determinado tópico em que as notícias mudam rapidamente e seria difícil acompanhar as mudanças por meio da leitura de todos os itens. Os procedimentos visam a produzir um resumo revisto a intervalos regulares (por exemplo, de hora em hora ou no início de cada dia). Cada resumo visa a mostrar apenas o que mudou desde o resumo precedente. Com a sumarização temporal, a 'novidade' torna-se um critério útil na seleção de frases. Isto é, uma frase nova, que seja bem diferente das frases que ocorreram no passado, é um candidato promissor para seleção. A seleção de frases que aparecerão nos resumos baseia-se numa combinação de novidade e 'utilidade' (probabilidade de relevância para o assunto). Esse tipo de rastreamento automático do desenvolvimento de uma notícia ao longo do tempo foi denominado 'rastreamento de eventos'. Ver, por exemplo, Yang et al. (2000).

Gong e Liu (2001) referem-se a resumos que se relacionam com determinado assunto como 'relevantes para a consulta'.

Mani (2001) faz uma revisão de trabalhos sobre várias formas de sumarização aplicada a apresentações de multimídia, que inclui tanto sumarização de áudio quanto de vídeo. Também é possível usar a sumarização automática junto com outros processos automáticos, como o emprego dos resumos produzidos como entrada para categorização de texto (ver, por exemplo, Kolcz et al., 2001).

À medida que os métodos de sumarização foram se tornando cada vez mais aprimorados, surgiam aplicações mais especializadas. Elas incluem sumarização de multidocumentos e miniaturização de textos.

A sumarização de textos não precisa restringir-se a um único texto. Na sumarização automática de multidocumentos (Mani, 2001), frases de muitas fontes independentes podem ser fundidas para formar um resumo. Por exemplo, todas as referências a determinada pessoa ou evento podem ser localizadas numa base de dados de textos, e essas referências comparadas para eliminar redundância e fundir o que restar em algumas frases proeminentes.

Schiffman et al. (2001) descrevem um sistema para criar um dossiê biográfico de uma pessoa mencionada com destaque nos noticiários mediante a extração de alusões presentes numa variedade de textos. "Selecionará e fundirá descrições de pessoas, extraídas de uma coleção de documentos, eliminando descrições redundantes." Efetua-se um alto nível de fusão e sumarização. Eis um exemplo (o texto sublinhado foi extraído diretamente das fontes; os conectivos, não sublinhados, são fornecidos pelo sistema):

> ***Henry Hyde is a Republican chairman of House Judiciary Committee and a prosecutor in Senate impeachment trial. He will lead the Judiciary Committee's impeachment review. Hyde urged his colleagues to heed their consciences, "the voice that whispers in our ear, 'duty, duty, duty'."****

Este resumo foi criado a partir de uma coleção de 1 300 itens de uma agência de notícias (707 000 palavras) que continham 503 frases que mencionavam Hyde.

Outra aplicação da sumarização a multidocumentos, descrita por Elhadad e McKeown (2001), refere-se a prontuários médicos. Os procedimentos que descrevem destinam-se a examinar as seções de resultados publicadas em artigos de periódicos recuperados numa busca, localizar o texto dos artigos que pareçam estar diretamente relacionados com as informações constantes do prontuário de um paciente e produzir um resumo relacionado a esse paciente.

O problema da sumarização de multidocumentos, bem como várias possíveis abordagens, é examinado por Goldstein et al. (2000) e por Mani (2001).

Atualmente muita atenção vem sendo dedicada à sumarização de textos (por

---

* Henry Hyde é presidente republicano da Comissão de Justiça da Câmara dos Deputados e promotor no processo de impeachment no Senado. Ele dirigirá a revisão do impeachment na Comissão de Justiça. Hyde conclamou seus colegas a ouvir suas consciências, "a voz que sussurra em seu ouvido, 'dever, dever, dever'." (N.T.)

exemplo, de mensagens de correio eletrônico) destinados a telas muito pequenas, como as de telefones celulares ou assistentes pessoais digitais. Corston-Oliver (2001) descreve um desses métodos ao qual denomina *compactação* de texto. As técnicas de compactação incluem seleção de frases, eliminação de caracteres e pontuação, e a substituição por abreviaturas de palavras por extenso ou expressões. Assim, uma frase como esta:

> *The problem of automatic summarization poses a variety of tough challenges in both NL understanding and generation.*

será compactada assim:

> *PrblmOfAutmtcSmmrznPssVrtyOfTghChllngsInBthNL UndrstndngAndGnrtn.*

Buyukkokten et al. (2001) descrevem métodos para sumarização de páginas da Rede para navegação com dispositivos portáteis de mão. Os procedimentos que desenvolveram incluem parcelamento de páginas da Rede em 'unidades textuais semânticas', que podem ser visualizadas completa ou parcialmente (por exemplo, somente a primeira ou as três primeiras linhas). O programa desenvolvido pode, alternativamente, identificar e exibir a) as palavras-chave mais importantes extraídas da unidade, b) a frase mais significativa, ou c) tanto as palavras-chave quanto a frase significativa. A seleção de palavra-chave baseia-se em número de ocorrências na unidade de texto e estimativas de ocorrência na Rede como um todo (com base na amostragem de 20 milhões de páginas). A seleção de frases usa versão modificada do método de Luhn para reconhecimento de frases significativas, já descrito neste capítulo. O desempenho relativo dos vários resumos foi avaliado com a participação de sujeitos humanos em tarefas de busca de informação. A combinação de palavra-chave e frase-chave foi a mais eficaz na conclusão da tarefa. Boguraev et al. (2001) também trataram da sumarização miniaturizada de notícias para dispositivos portáteis de mão.

O recente incremento de atividades em torno da sumarização de textos também acarretou um interesse renovado pelos métodos de avaliação (ver também o capítulo 9). Mani (2001) divide a avaliação de resumos preparados automaticamente em métodos intrínsecos e extrínsecos. Os métodos intrínsecos incluem:

a. utilizar um grupo de árbitros para decidir quais as frases que merecem ser selecionadas e as que não merecem (*concordância*)
b. avaliar a legibilidade do resumo em termos de certos critérios, como extensão da palavra e da frase e qualidade gramatical; para esse fim podem ser usados árbitros humanos ou corretores gramaticais e de estilo (*qualidade*)
c. comparar um resumo preparado automaticamente com um resumo 'ideal' preparado por seres humanos (*informatividade*)
d. avaliar um resumo em termos de se é capaz de responder um determinado conjunto de questões; o resumo pode ser comparado com o texto completo para esta avaliação (*método baseado em conteúdo*)
e. avaliar quanto da informação no texto completo é preservado no resumo (*fidelidade à fonte*).

Os métodos extrínsecos reconhecidos por Mani incluem 1) avaliar o resumo em termos de sua capacidade de prever corretamente a relevância do texto completo, 2) avaliar sua capacidade de permitir a um analista humano classificar corretamente o texto completo, e 3) avaliação da compreensão da leitura. Mani também reconhece o 'sistema avançado de avaliação', que envolve a avaliação de resumos no contexto de um sistema totalmente operacional (por exemplo, em termos de satisfação do usuário).

A sumarização de textos implica normalmente a extração de frases, embora sejam possíveis outros tipos de extração, como a de determinados termos ou tipos de termos, e talvez a colocação de termos extraídos em algum tipo de gabarito. Tomando um exemplo totalmente hipotético, um sistema poderia monitorar o movimento de executivos de empresas por meio da análise de notícias, e a frase

> **"João F. Cruzado, Vice-Presidente de Vendas da ABC durante os últimos cinco anos, foi nomeado Vice-Presidente Executivo da XYZ" seria reduzida à seguinte estrutura:**
>
> **Executivo: João F. Cruzado**
> **Cargo anterior: Vice-Presidente de Vendas**
> **Empregador anterior: ABC**
> **Novo cargo: Vice-Presidente Executivo**
> **Novo empregador: XYZ**
> **Data: 5 de novembro de 1996 (data da notícia)**

Cowie e Lehnert (1996) traçam um útil panorama sobre a extração de texto, e Grishman (1994) examina os problemas envolvidos na avaliação de resultados de trabalhos de extração. Shuldberg et al. (1993) oferecem a descrição minuciosa de uma abordagem. Onyshkevych (1994) e Hobbs e Israel (1994), entre outros, tecem considerações sobre o projeto de gabarito. Lawson et al. (1996) consideram este tipo de extração de dados/preenchimento de gabarito como uma forma de 'mineração de dados'. Esta denominação, no entanto, é aplicada com mais freqüência a procedimentos e programas que procuram descobrir nos dados (por exemplo, registros de vendas ou prontuários médicos) padrões e correlações significativas, sem instruções sobre o que procurar (ver capítulo anterior).

Há muitas aplicações potenciais para esse tipo de extração de texto e preenchimento de gabarito (quadro), das quais a mais óbvia talvez seja a produção de resumos de notícias atuais. Haug e Beesley (1992) examinam outra aplicação em que os dados de prontuários de pacientes podem ser reconhecidos automaticamente, extraídos e colocados sob um número limitado de cabeçalhos (por exemplo, 'queixas de', 'paciente nega') para ajudar os radiologistas na interpretação de radiografias. Paice e Jones (1993) examinam o emprego de uma abordagem de preenchimento de quadros na construção de resumos automáticos. Outra aplicação especializada do método de gabarito é a extração de citações bibliográficas do texto de patentes (Lawson et al., 1996). Humphreys et al. (2000) descrevem o modo como processos de preenchimento de gabarito podem ser aplicados à extração de determinados dados de periódicos científicos.

Os processos modernos de extração podem identificar textos candidatos (isto é, aqueles cujas palavras-chave indicam alta probabilidade de que o texto

conterá o tipo de dado a ser extraído) e porções do texto que sejam bons candidatos para os processos de extração, baseados numa combinação de análise sintática e semântica. Jacobs e Rau (1990) descrevem um desses sistemas aplicado à extração de informação sobre fusões de empresas. A extração de informação em geral é objeto de um livro organizado por Pazienza (1999).

Em algumas situações de recuperação, um conjunto limitado de características do texto pode ser da maior importância. Por exemplo, datas e nomes (de lugares, pessoas, organizações) são especialmente úteis em buscas de notícias. Watters e Wang (2000) descrevem um sistema capaz de extrair das notícias expressões substantivas próprias ['*name phrases*'] e categorizá-las (como local do evento, data do evento, nome pessoal, nome de instituição). O uso de iniciais maiúsculas é a deixa para identificação das expressões substantivas próprias. O sistema destina-se à recuperação interativa, em tempo real, baseada num algoritmo de comparação: o usuário com acesso em linha que encontra uma notícia de interesse pode pedir ao sistema que localize outras notícias que sejam semelhantes a essa. O sistema experimental pretende ser aplicado na Rede "usando como interface os navegadores comuns da Rede".

Vários processos foram desenvolvidos na National Library of Medicine (NLM) para identificação/extração em textos médicos. Bodenreider e Zweigenbaum (2000) lidam com a identificação de nomes próprios, Wilbur et al. (1999) com nomes químicos, Rindflesch et. al. (1999) com terminologia de ligação molecular, Rindflesch et al. (2000a) com terminologia de medicamentos e genes, e Sneiderman et al. (1998) com termos anatômicos. Em muitos casos, os termos identificados ou extraídos são ainda processados (por exemplo, para fazer a ligação com o Unified Medical Language System (UMLS). Outros projetos de pesquisa na NLM visam ao desenvolvimento de ferramentas lingüísticas para auxílio nesses tipos de processamento. Por exemplo, Weeber et al. (2001) lidam com o problema de desambiguação do sentido das palavras, e McCray et al. (2001) com o uso do sistema unificado de linguagem médica (UMLS) na identificação de expressões do texto que mereçam processamento adicional de linguagem natural.

Naturalmente, a sumarização extrativa não funciona bem com certos tipos de texto, inclusive páginas da Rede, que Berger e Mittal (2000) descrevem como "uma mixórdia caótica de expressões, vínculos, elementos gráficos e comandos de formatação". Descrevem seu trabalho no sentido de desenvolver 'sínteses essenciais' de páginas da Rede que não sejam extratos de texto (isto é, frases ou parágrafos), mas, sim, concatenações de palavras, como (um exemplo real) "the music business and industry artists raise awareness rock and jazz" [os artistas comércio e indústria música elevam consciência rock e jazz].

Jones e Paynter (2002) descrevem processos para extração automática de palavras-chave ou expressões-chave do texto dos documentos, com o objetivo de produzir sucedâneos que possam ser usados para pesquisar em extensas tarefas de recuperação de texto na Rede. A extração de 'expressões-chave' é feita por meio de processos de aprendizado de máquina. O algoritmo de extração

aprende com um conjunto de textos de treinamento nos quais as expressões-chave já foram atribuídas (por exemplo, por seus autores). Com base em avaliação feita por seres humanos, Jones e Paynter concluem que as expressões extraídas segundo seus procedimentos "não eram piores, estatisticamente, do que as apresentadas pelos autores". Anteriormente, Hui e Goh (1996) fizeram experiências com a geração automática de resumos de artigos de jornais como parte de uma interface de recuperação e filtragem da Rede.

Hoje em dia encontram-se disponíveis comercialmente programas para extração de várias formas de dados de sítios da Rede. Por exemplo, Ojala (2002) refere-se a um produto que poderá fazer buscas de mudanças na direção de empresas, compra e venda de empresas, resultados de reestruturação de empresas e outros indícios de mudança nessas organizações (entre outras aplicações).

As aplicações correlatas ao processamento de texto incluem vinculação de texto, aumento de texto e geração de texto.

A *vinculação de texto* emprega análises estatísticas e/ou sintáticas para identificar semelhanças entre diferentes passagens do texto, em geral de documentos completamente diferentes, e assim vinculá-los (Salton e Buckley, 1992; Maarek, 1992; Salton et al., 1997). Em essência, o método pode ser adotado para produzir automaticamente vínculos de hipertexto.*

A *ampliação de texto* pode ser considerada uma extensão da vinculação de texto. Os sistemas projetados para tal fim tentam integrar partes de textos de diversas fontes numa narrativa coerente — por exemplo, acompanhando notícias sobre um evento, como uma fusão de empresas ou um desastre natural, em jornais (e assim aplicável a tarefas de preenchimento de gabarito ou sumarização de multidocumentos). Variação disso é a pesquisa para desenvolver ferramentas que integrem entradas de textos e imagens — por exemplo, relacionar uma passagem descritiva num manual com elementos num diagrama e extrair texto que elucide o próprio diagrama (Rajagopalan, 1994). Chen (1993) descreve um computador 'modelo' para integração de textos afins oriundos de diferentes fontes.

Geração de texto refere-se a ferramentas de geração automática de tipos limitados de texto e a sistemas especialistas que ajudam as pessoas a redigir vários tipos de relatórios. Há sistemas deste tipo, por exemplo, que auxiliam na geração de documentação de produtos industriais, oferecendo acesso em linha a textos e elementos gráficos aplicáveis, de modo repetitivo, na criação de vários tipos de relatórios (ver Smith, 1991, por exemplo). Exemplo de sistemas mais elaborados é a 'bancada do editor', descrita por Bateman e Teich (1995), que poderá extrair texto e estruturá-lo em resposta a necessidades editoriais. 'Assistentes inteligentes de redação' modernos serão mais do que corretores ortográficos: orientarão na escolha de palavras, correção gramatical e uso do idioma (Oakman, 1994).

---

* Pozzi e Celentano (1993) analisam uma aplicação prática de vinculação, que inclui correspondência e outros documentos administrativos.

Stock (1993) e Stock et al. (1997) descrevem um interessante sistema de hipermídia (ALFRESCO) para recuperação de imagens de afrescos italianos do século XIV e informações a eles pertinentes. Entre outras características, o sistema incorpora uma interface de busca de linguagem natural e a capacidade de gerar respostas coerentes a partir de 'textos enlatados' relativos a diferentes imagens de afrescos armazenadas numa rede de hipermídia. Stock oferece um exemplo da pergunta 'Você poderia me mostrar e descrever um afresco de Ambrogio Lorenzetti em Siena?' que geraria a seguinte resposta:

> **Os Efeitos do Bom Governo é um afresco de Ambrogio Lorenzetti no Palazzo Publico. Os Efeitos do Bom Governo foi pintado em 1338. Um afresco do mesmo período é S. Silvestre e os Reis Magos, de Maso di Banco, pintado em 1330-1340. Outra obra de Ambrogio Lorenzetti num monumento de Siena é a Anunciação, de 1344, na Pinacoteca.**

Os vários componentes desta resposta foram extraídos de textos enlatados que aparecem em diferentes partes da rede de hipermídia.

Demasco e McCoy (1992) descreveram uma aplicação especializada de geração de texto. Seu trabalho visa a desenvolver uma interface que ajude pessoas que padeçam de graves deficiências motoras a compor textos. Um 'teclado virtual' permite ao usuário selecionar, em telas de letras, palavras ou expressões, e em seguida é usado um analisador [*parser*] semântico para gerar uma 'frase bem-construída'. Usam a denominação 'compansão de frase' [*sentence compansion**] para esse processo que poderia, por exemplo, tomar as palavras selecionadas 'João', 'estudo', 'meteorologia', 'grande' e 'universidade' e formar uma frase como 'João estuda meteorologia numa grande universidade'.

Kerpedjiev (1992) lida com outra situação especializada de geração de texto. Nela, são usados dados meteorológicos para gerar boletins meteorológicos 'multimodais'; os boletins podem ser em formato de texto narrativo, mapas, tabelas ou uma combinação das três formas, segundo as necessidades do usuário.

As tecnologias de recuperação da informação estão se difundindo numa ampla variedade de aplicações, onde antes eram pouco usadas, especialmente no mundo dos negócios. Com efeito, alguns dos métodos mais complexos de recuperação da informação foram mais bem aceitos por empresas comerciais do que pela indústria de serviços de informação. Uma importante aplicação é na área de atendimento a clientes [*help desk*].

Um serviço de atendimento a clientes é um serviço telefônico que lida com dúvidas e problemas dos clientes. Inicialmente, a denominação referia-se ao serviço mantido pela indústria de informática, a fim de lidar com problemas surgidos nas redes de computador. Embora os serviços de atendimento a clientes continuem muito ligados à indústria de informática, serviços similares existem hoje numa ampla variedade de empresas fabricantes de bens de consumo.

O coração de um típico serviço de atendimento a clientes é um 'componente de resolução de problemas', onde são armazenadas informações sobre problemas anteriormente encontrados e suas possíveis soluções. O serviço de atendimento a clientes pode ser considerado um verdadeiro sistema especialista: o pessoal de suporte que recebe as chamadas não é formado por especialistas, pelo menos não dominam todos os aspectos da situação, mas o sistema lhes oferece conhecimento para resolução de problemas. Estes serviços economizam no quantitativo de pessoal necessário para lidar com perguntas dos consumidores e reduzem a qualificação do pessoal designado para o serviço. Tornaram-se particularmente úteis porque muitos dos problemas ocorrem de modo repetitivo.

* Palavra formada a partir dos vocábulos *compression* e *expansion*. (N.T.)

Um bom exemplo de um serviço de atendimento a clientes, instalado na Compaq Computer Corporation, descrito por Acorn e Walden (1992), emprega uma versão do sistema de recuperação SMART desenvolvido por Salton. A utilização do sistema é exemplificada nas figuras 112 a 115. Os casos que foram tratados no passado (isto é, problemas e soluções) são armazenados na forma de descrições textuais, embora redigidos de modo sucinto e padronizado. O atendente que recebe a chamada de um cliente insere um enunciado textual do problema atual (figura 112). O sistema então procura casos semelhantes mediante busca em texto e apresenta 1) uma lista de casos de maior coincidência e 2) perguntas a serem feitas ao cliente, a fim de concentrar a busca e assim recuperar o caso e a solução correta. As respostas às questões afunilam o alcance da busca embora o consulente possa pesquisar informações sobre os casos armazenados (ver figura 113) para complementar as perguntas. Como resultado desse processo interativo, aos casos na base de dados são atribuídos escores numéricos que permitem que sejam ordenados por relevância provável. Um escore igual ou superior a 70 indica um caso que tem alta probabilidade de ser relevante.

A figura 114 mostra o exemplo de uma consulta, com perguntas geradas pelo sistema respondidas pelo cliente e os resultados apresentados como casos em ordem de relevância provável, e a figura 115 mostra o registro final do processo: o problema, as perguntas, o caso recuperado e a ação recomendada ao cliente. Os casos que não são resolvidos são analisados posteriormente por especialistas o que leva a novas adições à base de dados.

Os serviços de atendimento a clientes normalmente baseiam-se em interação que envolve o cliente, o representante do cliente e a base de dados. As perguntas geradas pelo sistema são necessárias para concentrar a busca com maior precisão. Em alguns casos, a resposta a uma pergunta genérica ('É um refrigerador *frost-free*?') pode restringir as ações seguintes a determinado segmento da base de dados (Danilewitz e Freiheit, 1991; Hart e Graham, 1997).

Os serviços de atendimento a clientes do tipo acima mencionado funcionam por meio de *raciocínio baseado em casos*. Embora o recurso de classificar em ordem de relevância provável não seja novidade, esses sistemas são inéditos pelo fato de que se concentram na solução mais provável mediante a geração de perguntas para o usuário extraídas dos próprios casos (por exemplo, A impressora foi instalada recentemente? Já tentou mudar o X? Já tentou limpar o Y?). Os

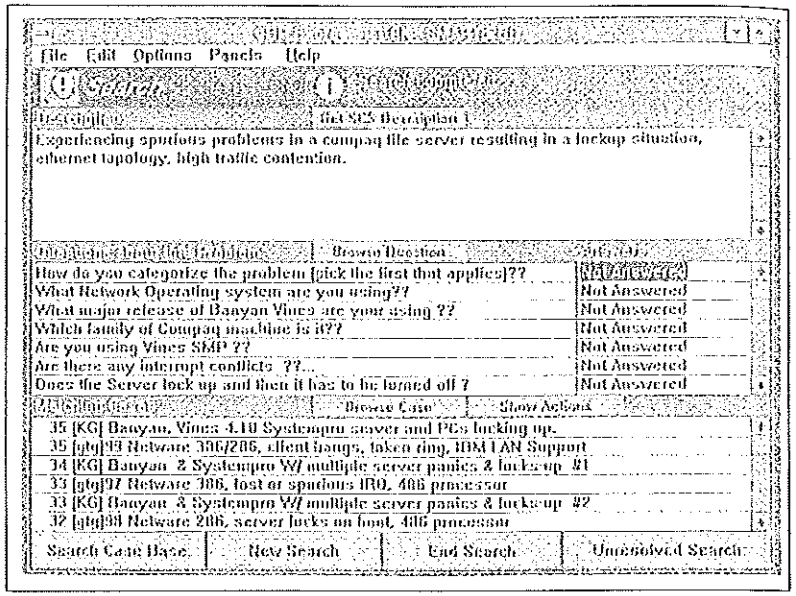


FIGURA 112

Busca inicial numa base de dados de um serviço de atendimento a clientes

Apud T.L. Acorn e S.H. Walden. In Scott, A.C.; Klahr, P., ed. *Innovative applications of artificial intelligence* 4, p. 3-18. Cambridge, MA, MIT Press, 1992.

casos neste tipo de base de dados podem ser construídos por 'autores de bases de casos' e existem no comércio programas que ajudam nessa tarefa.

Alguns serviços de atendimento a clientes incorporam métodos complexos de processamento de linguagem natural. Por exemplo, Anick (1993) descreve um desses sistemas, que também inclui uma forma de tesauro para ajudar os usuários a identificar termos de busca alternativos, e Uthurusamy et al. (1993) descrevem um sistema de diagnóstico que inclui processos altamente desenvolvidos para tornar mais inteligíveis descrições ambíguas ou malformuladas (registros de conserto de automóveis), graças à correção de erros ortográficos, desambiguação de abreviações e correção gramatical.

Os programas disponíveis no comércio para serviços de atendimento a clientes podem incorporar recursos de hipermídia, com texto, elementos gráficos, áudio e arquivos de vídeo acessíveis para ajudar no processo de diagnóstico. A integração de tecnologias de hipermídia e sistemas especialistas é revista por Ragusa e Turban (1994). Thé (1996) oferece um útil levantamento sobre programas para serviços de atendimento a clientes, disponíveis comercialmente no início de 1996.

Cada vez mais, as empresas que estão muito envolvidas em atividades de

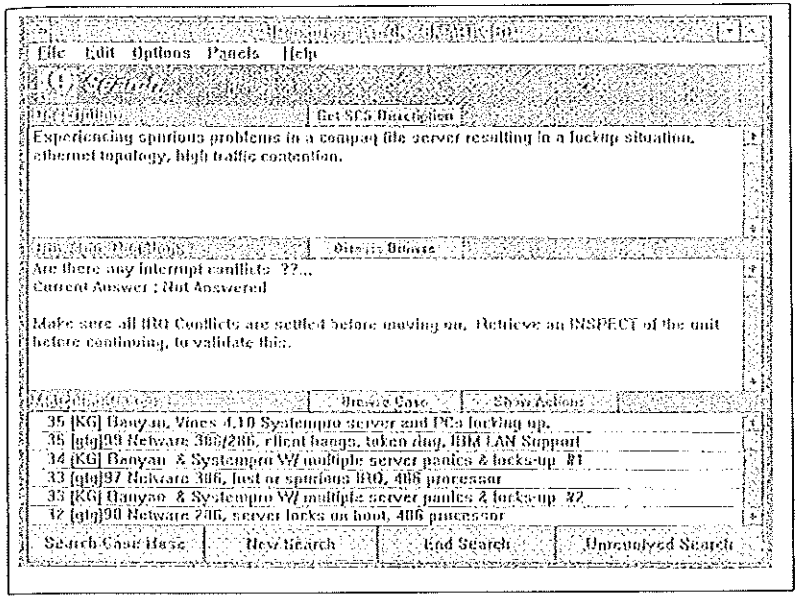


FIGURA 113

Pesquisa por mais informação em base de dados de serviço de atendimento a clientes

Apud T.L. Acorn e S.H. Walden. In Scott, A.C.; Klahr, P., ed. *Innovative applications of artificial intelligence* 4, p. 3-18. Cambridge, MA, MIT Press, 1992.

suporte a clientes vêm procurando desenvolver serviços que os próprios clientes possam usar para si, especialmente serviços que sejam implementados na Rede Mundial. Hoje em dia já existem programas, disponíveis comercialmente, que ajudam na implementação de serviços, baseados na Rede, de atendimento a clientes (Varney, 1996; Rapoza, 1996), que, inclusive, permitem aos clientes reportar problemas a um serviço de atendimento a partir de navegadores em suas estações de trabalho (Walsh, 1996).

Em livro de Moens (2000) encontra-se uma descrição bem completa dos processos examinados neste capítulo (e, em menor extensão, no anterior).

Conclusões

A recuperação da informação está implícita em todas as atividades de processamento de texto já mencionadas. Em termos de complexidade, a recuperação de frases ou parágrafos situa-se a meio caminho entre a recuperação de referências bibliográficas (típica da maioria das buscas em linha feitas em bibliotecas) e a recuperação de respostas reais a perguntas reais. Croft e Turtle (1992) asseguram que melhoramentos importantes na recuperação exigirão

técnicas que 'compreendam' o conteúdo de documentos e consultas e possam assim inferir se um item será útil.*.

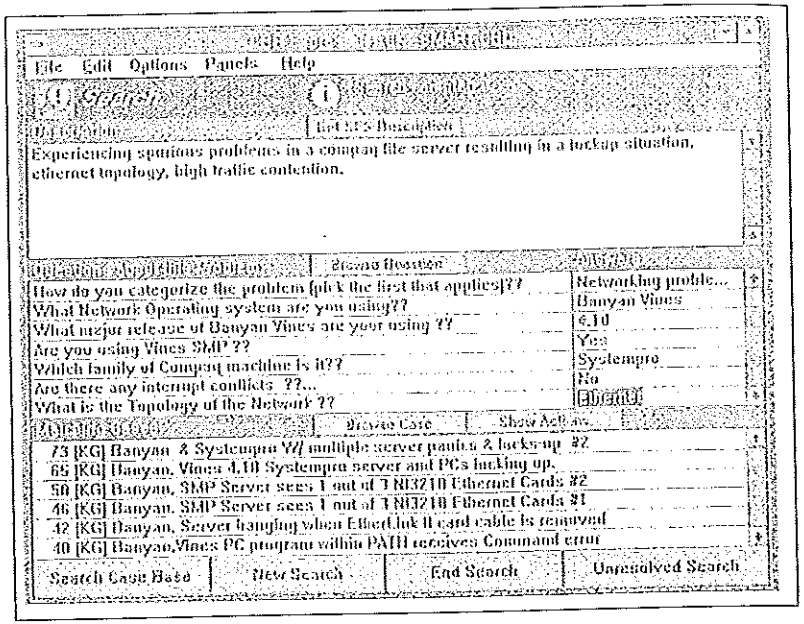


FIGURA 114

Casos com ordenação mais alta selecionados com base em consulta crítica e respostas dos clientes às perguntas

Apud T.L. Acorn e S.H. Walden. In Scott, A.C.; Klahr, P., ed. *Innovative applications of artificial intelligence* 4, p. 3-18. Cambridge, MA, MIT Press, 1992.

Os métodos hoje empregados em grande parte do processamento de texto não são particularmente novos. A maioria foi usada, talvez de modo mais rudimentar, há 30 anos, ou mais, por Luhn, Baxendale, Edmundson, Borko, Maron, Simmons, Salton e muitos outros pesquisadores (ver o capítulo 9 de Lancaster (1968b) para uma visão geral dessa área na década de 1960). Como foi sugerido, atualmente é possível alcançar melhores resultados porque há muito maior disponibilidade de conjuntos de textos eletrônicos e a potência dos computadores possibilita o processamento desses textos com razoável eficiência.

---

* Em algumas aplicações de processamento de texto o computador deve distinguir entre componentes lógicos do documento (por exemplo, título, resumo, texto principal, notas de rodapé, tabelas, figuras) e identificar relações entre eles (como, por exemplo, a ordem de leitura). Isso foi designado, um tanto pomposamente, 'compreensão do documento' (ver, por exemplo, Semeraro et al., 1994, e *Proceedings of the Third International Conference*, 1995).

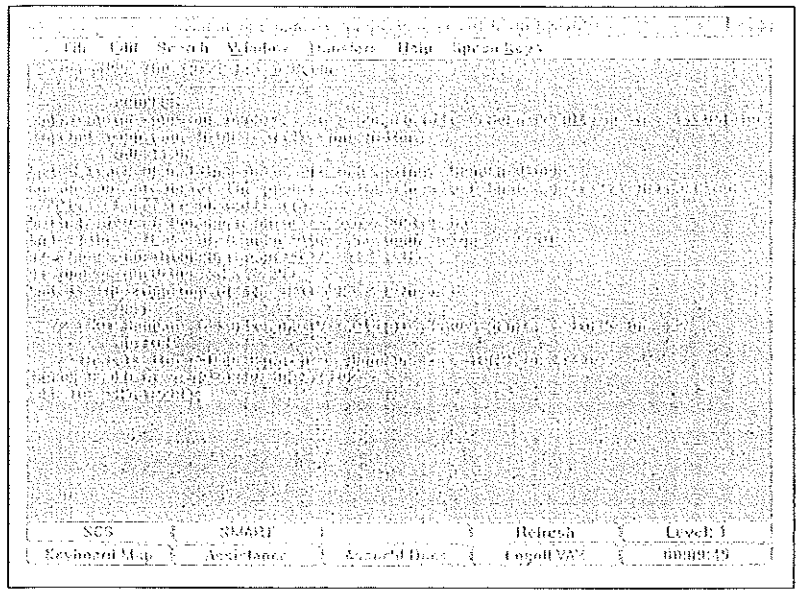


FIGURA 115

Resumo de caso com a ação recomendada ao cliente

Apud T.L. Acorn e S.H. Walden. In Scott, A.C.; Klahr, P., ed. *Innovative applications of artificial intelligence* 4, p. 3-18. Cambridge, MA, MIT Press, 1992

Não obstante, mesmo os métodos atuais mais complexos estão longe do ideal em termos de resultados alcançados, tempo e custos de processamento. Ademais, ainda são relativamente poucos os sistemas verdadeiramente 'operacionais' no sentido de que fornecem um serviço real de forma rotineira.

Jacobs (1992a) assim encara a situação:

> Embora hajam ocorrido alguns progressos visíveis na direção dos sistemas inteligentes baseados em textos, não chegamos muito perto de um estado aceitável de desenvolvimento da tecnologia (p. 5).

Hobbs et al. (1992) afirmam que o objetivo final é desenvolver um sistema que:

> [...] recuperará todas as informações que estejam, implícita ou explicitamente, presentes no texto, e concretizará isso sem cometer erros. Este modelo ainda está muito além do estado atual da tecnologia. Trata-se de uma meta incrivelmente alta para seres humanos, quanto mais para máquinas (p. 13-14).

McDonald (1992) ressalta que, em geral, os melhores analisadores [*parsers*] modernos somente lidam com frases relativamente curtas e simples. Com frases mais longas e complexas, o máximo que conseguem é identificar fragmentos componentes (por exemplo, um sintagma nominal); estão longe de produzir uma

análise completa e sem ambigüidade. No caso de uma frase de matéria jornalística, de tamanho comum, com 20 a 25 palavras, os analisadores atuais provavelmente chegariam a centenas de análises possíveis. Segundo McDonald, "nenhum analisador chega perto de compreender tudo num texto real, como uma reportagem".

Mesmo com *corpora* relativamente pequenos (cerca de 1 500 mensagens) de textos curtos (normalmente por volta de 14 frases), o melhor dos métodos atuais está longe de produzir resultados perfeitos — por exemplo, num exercício de extração de texto, nem todas as frases relevantes são selecionadas e nem todas as frases selecionadas são relevantes. Em condições controladas de avaliação, muitos sistemas modernos funcionam somente perto da marca de 50/50 (Jacobs e Rau, 1994; Sundheim, 1995) — por exemplo, produzem cerca de metade dos gabaritos (representações estruturadas baseadas em texto extraído das mensagens) que deveriam produzir e cerca de metade dos produzidos são esperados (isto é, coincidem com o modelo preestabelecido).* Embora alguns sistemas de processamento de texto reportem resultados muito melhores, isso ocorre com tarefas muito mais simples. Por exemplo, Hayes (1992a) relata 94% de revocação e 84% de precisão para o CONSTRUE, mas a tarefa executada — colocar notícias em até 200 categorias — é mais simples do que os trabalhos de extração de texto e preenchimento de gabarito.

Yang (1999) afirma que o CONSTRUE consegue alcançar tão bons resultados graças ao emprego de "regras, desenvolvidas manualmente, específicas de uma área ou específicas de uma aplicação" e que essa abordagem é muito cara para a maioria das aplicações.

Em condições controladas, é possível obter escores muito melhores em tarefas de extração mais simples (por exemplo, encontrar entidades nomeadas no texto) ou tarefas mais simples de preenchimento de gabarito — envolvendo extração de texto relativo a entidades nomeadas (Sundheim, 1995).

O nível de desempenho de 50/50 na extração de frase/conclusão de gabarito também precisa ser contextualizado. São resultados obtidos em áreas muito limitadas (por exemplo, atividade terrorista na América Latina). Para fazer a seleção de frases é preciso criar um dicionário específico da área. Mesmo numa área muito limitada, isso pode ser um trabalho que requer mão-de-obra intensiva (1 500 pessoas/hora foram mencionadas para um caso), embora hajam sido desenvolvidas ferramentas para construção desses dicionários automaticamente ou semi-automaticamente (ver Riloff e Lehnert, 1993, para um exemplo).**

---

* Vale a pena observar que os pesquisadores modernos da área de processamento de texto empregam as mesmas medidas — revocação e precisão — que foram descritas pela primeira vez na literatura de recuperação da informação na década de 1950.

** Uma ferramenta desse tipo 'aprende' com um *corpus* de texto de treinamento. Por exemplo, dado um conjunto representativo de extratos de textos que sabidamente tratam do tópico *x*, ela construirá um dicionário capaz de selecionar, a partir de um novo *corpus* de textos, frases sobre o tópico *x*.

A tarefa de encaminhamento na TREC-6 (1997) conseguiu, quando muito, apenas 42% de precisão (Voorhees e Harman, 2000) com somente 47 assuntos.

Em geral, mesmo os mais complexos dos atuais processos de indexação automática saem perdendo na comparação com a indexação feita por seres humanos qualificados. Por exemplo, Chute e Yang (1993), trabalhando com relatos de casos cirúrgicos, constataram que os códigos de procedimento atribuídos por seres humanos produziam melhores resultados do que vários processos automáticos, inclusive a indexação semântica latente. Anteriormente, Hersh e Hickam (1991) relataram que buscas em palavras do texto (somente títulos e resumos) davam melhores resultados do que buscas em registros indexados por pessoas (MEDLINE) ou processados automaticamente em ambiente médico. Mais tarde (Hersh e Hickam, 1995a) relataram 'nenhuma diferença significativa' em buscas num manual médico com dois métodos de processamento automático (um baseado em palavras e outro 'baseado em conceitos') e uma abordagem booleana de buscas em texto. Hersh e Hickam (1995b) fazem um apanhado dos estudos de avaliação que realizaram durante um período de quatro anos.

Moens e Dumortier (2000) descrevem um método de atribuição de categorias a artigos de revistas de interesse geral. O índice de êxito relatado é muito modesto. Com apenas 14 categorias genéricas para atribuir, o máximo que seus procedimentos conseguem alcançar não passa de 74% de revocação e 64% de precisão. Isto é, esse método atribuía 74% das categorias que deviam ser atribuídas enquanto 64% das atribuições realmente feitas foram consideradas corretas.

Para Fidel (1994) a indexação automática se orienta para os documentos ao invés de centrar-se nos usuários. Embora isso seja, em geral, verdadeiro, é possível fazer métodos mais centrados nos usuários, como, por exemplo, o uso de listas de termos a serem procurados especificamente num texto. Como a própria Fidel também realça, um sistema totalmente automatizado pode ser mais centrado nos usuários no lado da saída, ao permitir consultas em linguagem natural, retroalimentação de relevância e saída em ordem de provável relevância.

Ademais, muitos dos sistemas automáticos utilizam uma forma de ponderação para produzir uma saída em ordem de provável relevância. Embora alguns estudos (por exemplo, Salton, 1972) hajam reivindicado êxito na ordenação por relevância provável, outros não o fizeram. Num estudo de recuperação de informação, Marchionini et al. (1994) obtiveram resultados muito inferiores na ordenação por relevância provável. Isso também aconteceu em aplicações completamente diferentes. Por exemplo, sistemas de diagnóstico automático em medicina raramente colocam o diagnóstico 'correto' no alto da ordenação e com freqüência ele aparece bem lá embaixo (Berner et al., 1994; Kassirer, 1994).

Abordagens mais modernas para produzir resumos 'inteligentes' (resumos automáticos) de documentos não chegam a impressionar. O sistema desenvolvido e avaliado por Brandow et al. (1995) produziu resumos que foram julgados significativamente menos aceitáveis do que o 'lide do texto'. O que isso significa é que analistas humanos, em média, julgam as primeiras 250 palavras (suponha-

mos) de um texto como um indicador de conteúdo melhor do que um resumo de 250 palavras formado com frases selecionadas do texto automaticamente.

Moens (2000), que trabalhou durante algum tempo na área de sumarização automática, concorda que um resumo gerado automaticamente é apenas "uma aproximação de um resumo ideal".

Gaizauskas e Wilks (1998), depois de uma excelente revisão do campo da extração de texto (referem-se a ela como extração de informação (EI)), concluem ser improvável que os níveis de desempenho característicos dos sistemas de recuperação de informação se prestem à maioria dos propósitos:

> Os escores combinados de precisão e revocação para os sistemas de recuperação da informação [RI] mantiveram-se na faixa intermediária dos 50% durante muitos anos, e é nessa faixa que hoje se encontram os sistemas de EI. Embora os usuários de sistemas de RI tenham se adaptado a esses níveis de desempenho, não está claro que sejam aceitáveis para as aplicações de EI. É claro que o tolerável variará de uma aplicação para outra. Mas quando as aplicações de EI envolvem a construção de bases de dados que cobrem longos períodos de tempo, que subseqüentemente formam a entrada para análise ulterior, o ruído nos dados comprometerá seriamente sua utilidade (p. 97).

O interesse grandemente renovado por processos automáticos aplicados a vários aspectos da recuperação da informação atraiu muitos grupos de pesquisa para este campo pela primeira vez. Carentes de perspectiva histórica, é provável que dupliquem trabalhos feitos no passado ou, no mínimo, deixem de se fundamentar em pesquisas anteriores. Só um exemplo: Fowler et al. (1996) e Zizi (1996) descrevem trabalho sobre visualização gráfica de conceitos em linha que é muito parecido com o trabalho de Doyle (1961) realizado 40 anos antes.

No capítulo anterior, foi mostrado que muitas pesquisas que comparam a recuperação de texto com a recuperação de bases de dados indexadas padecem de sérias falhas. Infelizmente, o mesmo se pode dizer sobre comparações entre processos de indexação automática e indexação por seres humanos. Há um exemplo em Hmeidi et al. (1997). Com base em resultados de recuperação numa pequena base de resumos em árabe no campo da ciência da computação, os autores concluem que a "indexação automática é pelo menos tão eficaz quanto a indexação manual e mais eficaz em alguns casos". Na realidade, não houve indexação manual: a comparação foi entre um processo de indexação automática baseado no trabalho de Salton e busca em texto aplicada aos resumos. Uma vez que os procedimentos *à la* Salton destinam-se a melhorar tanto a revocação quanto a precisão (por exemplo, com a eliminação de palavras de maior e menor freqüência), é natural que tenham alcançado melhores resultados do que se nada fizessem com o texto além talvez de reduzi-lo à forma de tema/raiz.

Apesar de os custos de computação continuarem caindo, o processamento atual de textos não é necessariamente uma proposta barata. Hayes (1992a) coloca isso em perspectiva, com o exemplo do CONSTRUE. Em 1992, o sistema processava texto à velocidade de cerca de 1 800 palavras por minuto (uma mensagem média da Reuters, com 151 palavras, em menos de cinco segundos). A essa velocidade, ele alerta, "um gigabyte de texto tomaria quase dois meses de tempo de CPU para categorizar". Isso se refere à mera colocação de notícias em cerca de 200 categorias. É claro que a extração e manipulação de texto, que são mais complexas, tomariam mais tempo de processamento. Hayes, em tom dramático, salienta que o CONSTRUE ocuparia 20 anos ou mais de tempo de CPU para processar uma base de dados de 100 gigabytes do tamanho de NEXIS. Embora válido, há um equívoco nisso. Um sistema como o CONSTRUE foi projetado para *corpora* relativamente pequenos, um de cada vez — por exemplo, todas as mensagens recebidas num único dia — e não para análise de gigantescas bases de dados retrospectivas. Isso lembra a distinção entre buscas retrospectivas em grandes bases de dados bibliográficos na década de 1960, mediante processamento em lotes, e o uso de atualizações das mesmas bases para notificação corrente (DSI). Esta era economicamente atraente; as buscas retrospectivas certamente não.

O desenvolvimento dos atuais sistemas de uso limitado é também muito caro. CONSTRUE, por exemplo, exigiu 9,5 pessoa/anos de esforço (Hayes e Weinstein, 1991).

É interessante observar que o tipo de resultados 50/50 reportado por alguns sistemas modernos de processamento de textos está muito perto do nível de desempenho relatado para grandes sistemas de recuperação bibliográfica (por exemplo, MEDLARS) na década de 1960 (Lancaster, 1968a). Embora, na superfície, a comparação pareça injusta, posto que as tarefas de extração de texto e preenchimento de gabaritos são claramente mais complexas do que o trabalho de recuperação de referências, deve-se também reconhecer que os *corpora* usados nas tarefas mais complexas são bastante pequenos em comparação com o tamanho das bases de dados bibliográficos, mesmo as de 30 anos atrás.*

O fato é que os métodos booleanos de busca, relativamente toscos, mais comumente usados para pesquisar em grandes bases de dados bibliográficos hoje em dia, apesar de seus inúmeros críticos, apresentam resultados notavelmente bons, considerando o tamanho dos *corpora* com que lidam, um ponto levantado de modo muito convincente por Stanfill e Waltz (1992):

> O que surpreende (do ponto de vista da IA) é que a abordagem estatística, que não utiliza absolutamente qualquer conhecimento específico de uma área, funciona. E funciona com quantidades de informação (gigabytes) que são incrivelmente grandes pelos padrões da IA (p. 217).

Note-se que estavam se referindo aos simples métodos booleanos de busca usados em bases de dados indexadas (por exemplo, MEDLINE) ou de texto completo (por exemplo, NEXIS) e não às abordagens mais complexas de saídas ordenadas por relevância provável.

---

* Em exercícios mais convencionais de recuperação, os métodos modernos de buscas em texto nem mesmo alcançam o nível 50/50 de desempenho quando estão envolvidas bases de dados muito maiores (centenas de milhares de itens) (Harman, 1997; Sparck Jones, 1995).

Jacobs (1992a) identificou vários desafios que se colocam hoje para os pesquisadores na área de processamento de texto: tornar os sistemas mais robustos (maior exatidão, mais rápidos, mais baratos na análise lingüística), recursos de refinamento (por exemplo, passar de recuperação de documento para recuperação de trechos para recuperação de resposta), e fazer saídas com melhor relação custo–eficácia ou atraentes para o usuário (mediante realçamento, extração de texto ou sumarização).

Embora algum progresso haja certamente ocorrido na aplicação do computador a várias tarefas relativas à recuperação de informação, existem reduzidos indícios de que os processos automáticos possam ainda vir a superar os seres humanos em tarefas tão intelectuais como indexação, redação de resumos, construção de tesauros e criação de estratégias de busca.

Kuhlen (1984) sugere que ainda não conhecemos o suficiente sobre os processos intelectuais envolvidos na redação de resumos (e, por analogia, na indexação e procedimentos afins) para desenvolver programas com os quais essas atividades seriam simuladas pelo computador:

> Resumir [...] é uma arte intelectual e, como tal, não transferível diretamente para processos automáticos. A psicologia cognitiva e a inteligência artificial ainda não nos propiciaram conhecimento suficiente sobre os processos que se passam de fato nas mentes dos resumidores quando compreendem textos e os condensam. Assim, a imitação direta de um processo intelectual, como o ato de resumir, parece inalcançável (p. 98).*

Apesar da pesquisa e desenvolvimento verificados mais ou menos na última década, as palavras de Kuhlen parecem até hoje pertinentes.

* Alguns autores cometem o erro de afirmar que os tipos de processamento automático de texto analisados neste capítulo (por exemplo, atividades de extração) constituem 'compreensão automática' de texto (ver, por exemplo, Moens et al. (1999). Nada estaria tão longe da verdade.

Capítulo 16

A indexação e a internet

A Rede Mundial tornou-se tão imensa, desajeitada e complexa nos últimos anos que se distanciaria muito do objetivo deste livro tentar explicar a míriade de componentes que formam seu aparato: navegadores, mecanismos de busca, mecanismos de metabusca, agentes de busca, *crawlers*, etc. O panorama descrito por Schwarz (1998) ainda é conceitualmente útil, embora muitas de suas informações estejam desatualizadas. Encontra-se em Arasu et al. (2001) uma excelente descrição técnica sobre como realmente funcionam os *crawlers* e mecanismos de busca da Rede. Em Liddy (2002) encontra-se uma abordagem mais legível (isto é, mais simples). As mudanças ocorrem agora tão rapidamente que qualquer relato, mal é publicado já está, pelo menos em parte, obsoleto. Para se manter a par das mudanças, é preciso usar um serviço como

<http://extremesearcher.com/news.htm>

e

<http://searchenginewatch.com>

encontrados na própria Rede. Este capítulo limitar-se-á a questões da Rede que sejam mais pertinentes aos temas da indexação e redação de resumos.

Da perspectiva da recuperação, os recursos informacionais acessíveis na Rede são muitíssimo diferentes dos registros bibliográficos do sistema 'convencional' da figura 1 no começo do livro. Entretanto, há certas semelhanças. Os sítios da Rede têm diferentes elementos pesquisáveis: URL, nome do sítio, quaisquer metadados incluídos no sítio, páginas de conteúdo (onde os termos que ali ocorrem podem ser considerados um tanto análogos aos termos de indexação de registros bibliográficos convencionais) e todo o texto encontrado no sítio.

Esses elementos pesquisáveis apresentam de fato algumas semelhanças com os diferentes elementos pesquisáveis em registros presentes numa base de dados bibliográficos: título, números de classificação (às vezes), termos de indexação, texto de resumo (embora a quantidade de texto no sítio da Rede possa ser consideravelmente maior). Os sítios da Rede diferem da maioria dos registros bibliográficos pelo fato de que podem também conter apontadores (vínculos de hipertexto) para outros sítios, onde os termos dos vínculos são também pesquisáveis. Neste sentido, um sítio não é uma unidade independente (como um registro bibliográfico comum), mas um nó de uma rede.

Lynch (2001) chamou atenção para uma diferença importante entre uma base de dados bibliográficos, como o MEDLINE, e o conjunto de sítios que formam a Rede Mundial. As organizações produtoras de bases de dados bibliográ-

ficos são 'neutras' em face dos registros que processam. Os sucedâneos que desenvolvem — resumos e termos de indexação — destinam-se a representar de modo exato e imparcial os documentos. Por outro lado, muitos sítios foram criados por empresas que almejam que sua página seja recuperada e não a produzida pelo concorrente. Há dois modos de fazer isso: o '*index spamming*' [saturação de índice] e o '*page jacking*' [seqüestro de página] (ver adiante).

Serviços de busca na Rede

Hock (2001) nos dá uma clara visão global da Rede como um sistema de recuperação de informação:

> Para o que nos interessa, um mecanismo de busca é um serviço oferecido por intermédio da Rede Mundial que permite ao usuário dar entrada a uma consulta e fazer buscas numa base de dados que abrange uma porção bastante substancial do conteúdo da Rede. Para ser um pouco mais específico, um mecanismo de busca permite ao usuário ingressar com um ou mais termos, e qualificadores opcionais, a fim de localizar páginas de interesse na Rede. O termo é quase intercambiável com 'serviços de busca na Rede', que [...] normalmente se refere mais ao sítio como um todo, e que por sua vez pode fornecer o mecanismo de busca como uma de múltiplas opções. O mecanismo de busca pode até ser apenas uma oferta num conjunto de ofertas que juntas visam a oferecer ao usuário um lugar geral de partida ou um 'portal' da Rede (p. xxii).
>
> Pode-se visualizar o próprio mecanismo de busca como se fosse composto de cinco partes principais funcionais: 1) os '*crawlers*' do mecanismo, que saem em busca de sítios e páginas da Rede; 2) a base de dados de informações reunidas sobre essas e outras páginas que hajam sido reunidas a partir de outras fontes; 3) o programa de indexação, que indexa o conteúdo da base de dados; 4) o 'mecanismo de recuperação', o algoritmo e programação respectiva, dispositivos, etc., que, a pedido, recuperem material do índice/base de dados; 5) a interface gráfica (HTML), que reúne dados da consulta do usuário para alimentar o mecanismo de recuperação (p. 6).
>
> *Crawlers* ou *spiders* são programas que percorrem a Rede para 1) identificar novos sítios que serão acrescentados ao mecanismo de busca e 2) identificar sítios já cobertos, mas que tenham sofrido mudanças. Os *crawlers* coletam informações sobre o conteúdo das páginas de sítios e alimentam a base de dados do mecanismo de busca com essas informações (p. 6).
>
> O conjunto total de informações armazenadas sobre todas as páginas da Rede constitui a base de dados do mecanismo de busca. Esse conjunto inclui páginas identificadas por *crawlers*, mas, cada vez mais, inclui também páginas identificadas por outras fontes ou técnicas. Um número muito grande de sítios acrescentados aos mecanismos de busca tem origem em solicitações feitas diretamente pelos editores de páginas da Rede (p. 7).
>
> Em termos de quais páginas serão realmente recuperadas por uma consulta, a indexação pode até ser mais crítica do que o processo dos *crawlers*. O programa de indexação examina as informações armazenadas na base de dados e cria as entradas apropriadas no índice. Quando se submete uma consulta, é esse índice que é usado a fim de identificar registros coincidentes.

> A maioria dos mecanismos de busca afirma que indexa 'todas' as palavras de toda página. O logro está no que os mecanismos escolhem para considerar 'palavra'. Alguns trabalham com uma lista de 'palavras proibidas' [...] que não são indexadas (p. 8).
>
> Todos os principais mecanismos indexam os campos de 'alto valor', como título e URL. Comumente, mas nem sempre, indexam-se as metaetiquetas [*metatags*], que são palavras, expressões ou frases colocadas numa parte especial do código HTML (Hypertext Markup Language) como forma de descrever o conteúdo da página. As metaetiquetas não aparecem ao se visualizar uma página, embora se possa vê-las, caso se queira, pedindo ao navegador para mostrar a 'página-fonte'. [...] Alguns mecanismos, porém, propositalmente, *não* indexam algumas metaetiquetas porque eles são a parte da página que é mais suscetível de violação pelos *spammers*. Esta cautela é adotada em detrimento da indexação de informações extremamente úteis (p. 8-9)

Schwarz (1998) abordou a variabilidade da indexação nos vários serviços:

> Alguns serviços indexam todas as palavras de uma página [...] Informações posicionais e informações de etiquetas de marcação podem ser armazenadas com texto indexado para melhorar a recuperação e a eficácia da ordenação. Outros somente indexam palavras que ocorram freqüentemente, ou apenas palavras que ocorram dentro de certas etiquetas de marcação, ou só as primeiras *n* palavras ou linhas [...] Poderão ou não ser adotadas listas de palavras proibidas, e, se forem adotadas, poderão incluir palavras que ocorrem com freqüência muito alta [...] (p. 975).

O estudo de Hert et al. (2000) é um dos poucos que analisam um sítio da Rede de uma perspectiva de indexação. Os pesquisadores desenvolveram três abordagens alternativas para a indexação realmente utilizada num sítio existente e as compararam por meio de buscas feitas por 20 estudantes universitários. As comparações foram feitas com base tanto na eficácia da recuperação quanto das preferências dos usuários.

Embora vários mecanismos de busca aleguem possuir atualmente bases de dados com mais de 200 milhões de registros, nenhum desses mecanismos consegue abarcar todos os sítios da Rede. Lawrence e Giles (1999) estimaram que a cobertura não passava de 16% no caso do mecanismo mais exaustivo, e muitos cobriam apenas 10% ou menos. Além disso, relataram que a cobertura parecia estar diminuindo com o passar dos anos. Quer dizer, a Rede estava crescendo numa velocidade mais rápida do que os mecanismos de busca podiam suportar.

Recursos de recuperação

Embora o usuário comum da internet provavelmente faça suas buscas inserindo uma seqüência simples de termos (que diferentes mecanismos de busca tratarão de modo diferente) — alguns colocando os termos numa relação OU, outros numa relação E), esses mecanismos podem na realidade oferecer várias opções mais avançadas — como o emprego de:

1. Lógica booleana, inclusive recursos de encaixamento [*nesting*]
2. Truncamento

3. Buscas com expressões
4. Proximidade de palavras
5. Buscas em campos (isto é, poder limitar a busca a um campo especificado no registro, como título ou URL)
6. Vínculos de hipertexto (isto é, buscar páginas vinculadas a determinado URL)
7. Busca de imagens (capacidade de procurar apenas páginas que contenham imagens)
8. Consulta por exemplo (capacidade de encontrar registros semelhantes a um registro já conhecido como interessante).

Naturalmente, nem todos os mecanismos de busca possuem todos esses recursos.

Os registros recuperados numa busca na Rede são ordenados com base num escore numérico e apresentados ao usuário nessa ordem. Vários fatores podem ser levados em conta nessa pontuação, inclusive:

1. Freqüência de ocorrência de termos de busca no registro. Pode ser usada a freqüência relativa (o número de ocorrências é relacionado com a extensão do registro, de modo que, por exemplo, um termo de busca que ocorra cinco vezes num registro de 100 palavras terá peso maior do que outro termo que ocorra cinco vezes num registro de 1 000 palavras). É provável haver um limite para o número de ocorrências levadas em consideração na pontuação devido ao *index spamming*.
2. Número de coincidências de termos. Registros que coincidem com todos os três termos numa consulta (por exemplo) alcançam pontuação maior do que os que coincidem com apenas dois.
3. Localização do termo. Termos que ocorrem no título podem obter mais peso do que aqueles que ocorrem em outros lugares.
4. Raridade. Termos muito incomuns — os que ocorrem muito poucas vezes na base de dados — têm probabilidade de alcançar escore mais elevado.
5. Proximidade. Se os termos de busca ocorrerem muito próximos no texto, isso pode valer mais do que se estivessem muito distantes um do outro.
6. Ordem dos termos. Um termo que haja sido inserido em primeiro lugar pelo consulente pode receber um peso maior do que os subseqüentes.
7. Data. Os registros mais recentes obtêm peso maior.
8. Popularidade da fonte baseada ou no número de vezes que foi acessada ou no número de outras fontes a ela vinculadas.

Alguns mecanismos de busca também permitem formatos de saída alternativos — essencialmente opções de visualização sucintas *versus* extensas.

Eastman (2002) descobriu que itens que apareciam no topo da classificação no curso de buscas relativamente simples (uma seqüência de termos de busca) possuíam maior relevância (precisão) do que aqueles que apareciam no topo da classificação quando se empregavam buscas booleanas mais complexas para os mesmos assuntos. Ela conclui que os processos de classificação do mecanismo de busca devem funcionar muito bem.

Hock (2001), no entanto, assegura que as buscas na Rede ainda são muito rudimentares se comparadas com o uso de uma base de dados cuidadosamente indexada, como o MEDLINE, por parte de um consulente experiente. Reconhece, porém, que os recursos de busca na Rede vêm melhorando com o tempo:

> O hiato entre as expectativas da recuperação tradicional e as expectativas das buscas na Rede diminui ainda mais quando se levam em conta dois outros fatores. O reconhecimento de ambos os fatores é importante para o consulente que queira tirar o máximo de proveito de qualquer um desses tipos de serviço de busca.
>
> Primeiro, os mecanismos de busca na Rede lidam com dados muito desestruturados, ou pelo menos dados com uma estrutura muito pouco coerente. De fato, existe uma estrutura definida para a HTML por trás das páginas da Rede, mas, no que diz respeito ao real conteúdo intelectual, a quase única estrutura 'intelectual' encontra-se nos títulos e metaetiquetas. O corpo das páginas tem pouca estrutura coerente que o serviço de busca na Rede possa usar em buscas estruturadas [...]
>
> Segundo, o simples volume de dados atualmente na Rede — associado ao volume que aumenta todos os dias — acrescenta um grau de respeito ao que os mecanismos de busca na Rede conseguiram realizar em período de tempo muito curto. O fato de existir pelo menos um nível elementar de acesso às centenas de milhões de páginas de material é um feito que deve inspirar mais admiração do que frustração (p. 20-21).

Outros autores são mais críticos acerca dos recursos de busca na Rede. Wheatley e Armstrong (1997), por exemplo, expõem assim a situação:

> No corpo de uma página da Rede, não há a possibilidade de dados em campos definidos, e por isso é impossível [...] limitar as buscas a [...] partes da página. Portanto, uma busca descobrirá o(s) termo(s) de busca com igual facilidade no último parágrafo, em nota de rodapé explicativa ou em material existente perto do alto da página. Com o advento dos metadados, torna-se [...] [possível] uma abordagem levemente diferente [...] Mas, como os metadados não são exibidos e em geral inexiste editora ou autoridade que imponha limites, é fácil ocorrer abuso de palavras-chave e pares de termos descritivos e recheados de termos destinados a dar-lhes alta relevância aparente ou localização freqüente [...] Inexiste, até hoje, norma para atribuição de rótulos a recursos em rede, e, embora continuem os estudos sobre metadados, seu uso efetivo na área superior de páginas da Rede ainda é raro e incoerente (p. 206).

Entretanto, desde que isso aí em cima foi escrito, em certa medida as coisas melhoraram, tanto na questão dos campos quanto dos metadados.

Naturalmente, diferentes mecanismos de buscas produzirão resultados diferentes para a mesma consulta por causa de diferenças de cobertura, de algoritmos de busca e critérios de ordenação. Muitas comparações de desempenho surgiram na literatura nos últimos anos, remontando a 1995, mas são de valor limitado devido à situação de constante mudança da própria Rede.

Comparações avaliatórias cotejam resultados de buscas somente com base na duplicação/unicidade ou tentam estabelecer a relevância dos itens recuperados. Leighton e Srivastava (2000) e Su e Chen (1999) são exemplos do último caso. Jansen e Pooch (2001) fazem uma revisão de estudos anteriores. Outras

avaliações foram feitas com objetivos especiais. Thelwall (2001) compara os mecanismos de busca em relação a seu emprego potencial em aplicações de mineração de dados, que ele parece definir como "a agregação de informações oriundas de grande número de páginas da Rede, para criar conhecimento novo".

Oppenheim et al. (2000) fazem excelente revisão de avaliações de mecanismos de busca feitas anteriormente. Recomendam o desenvolvimento de um conjunto normalizado de procedimentos para essas avaliações, de modo que "sejam feitas comparações de mecanismos de busca de modo mais eficaz, e que sejam rastreadas as variações de desempenho de qualquer mecanismo de busca ao longo do tempo".

Em certa medida, é possível compensar os recursos entre os mecanismos de busca por meio do emprego de mecanismos de metabusca, serviços que fazem buscas em vários mecanismos de busca, e em seguida agrupam os resultados. Segundo Hock (2001), existem hoje mais de 100 mecanismos de metabusca em uso. Hock é claro quanto às suas limitações:

> Em particular, se houver mais de um punhado de sítios relevantes para encontrar nos mecanismos de busca, os mecanismos de metabusca freqüentemente não encontrarão a maioria deles. Isso é causado por vários fatores, inclusive os limites impostos pelo serviço ao número de registros recuperados em cada mecanismo, limites de tempo quando o serviço de metabusca simplesmente interrompe a busca num mecanismo se demorar muito, incapacidade de traduzir adequadamente a consulta para a sintaxe específica exigida pelo mecanismo-alvo, e outros fatores. Felizmente, alguns mecanismos de metabusca realmente conseguem captar todos os registros que ali existam (mas têm outros inconvenientes).
>
> Os três principais pontos fracos dos mecanismos de metabusca são: 1) muitas vezes limitam estritamente o número de registros que recuperarão de cada mecanismo (às vezes não mais de dez); 2) muitas vezes não repassam aos mecanismos consultas que tenham um mínimo de complexidade; e 3) na maioria dos casos, só fazem buscas em dois ou três dos maiores mecanismos de busca [...]
>
> Na maioria, os mecanismos de metabusca diferem entre si nos seguintes aspectos:
>
> - Os mecanismos de busca específicos que abrangem
> - O número de mecanismos de busca que podem ser pesquisados por vez
> - A capacidade de repassar consultas mais complexas — como as que incluem expressões, enunciados booleanos, etc. — para os mecanismos de busca 'alvos'
> - Limites quanto ao número de registros que podem recuperar de cada mecanismo (que pode chegar a ser no máximo 10)
> - O tempo que estão dispostos a gastar na busca em cada mecanismo (antes de interromper a sessão por decurso de tempo)
> - Como a saída é apresentada, inclusive se eliminaram ou não registros certos encontrados em duplicata nos vários mecanismos (p. 186-187).

Hock salienta que os mecanismos de metabusca são mais úteis quando se procura algo obscuro, isto é, assuntos sobre os quais provavelmente há muito poucos sítios que tenham algo a oferecer.

No início do capítulo, fez-se uma distinção entre os recursos da Rede e as

bases de dados bibliográficos tradicionais. Naturalmente, muitas dessas bases de dados estão disponíveis na Rede. Embora uma base de dados como o MEDLINE possa ser localizada e nela serem feitas buscas, seu conteúdo não é incluído nos resultados apresentados por nenhum dos mecanismos de busca, aspecto que foi explicitado por Zich (1998):

> A busca de informações na Rede padece de duas deficiências debilitantes — o processo de buscas é superficial e limitado. É superficial porque os mecanismos de busca chegam somente até ao que chamo de documentos de primeiro nível — isto é, documentos que residem em servidores em HTML. Há um mundo de informações adicionais além desse ponto. Refiro-me a informações em catálogos de bibliotecas e outros arquivos de dados a que a Rede oferece acesso. O catálogo da Library of Congress, por exemplo, nunca é pesquisado por nenhum dos mecanismos de buscas. Milhões de informações, meticulosamente organizadas e rigorosamente autenticadas, que ali jazem e estão disponíveis continuam inexplorados por esses mecanismos. As descrições dos materiais em formato digital do programa American Memory, da Library of Congress, e os próprios materiais digitalizados — centenas de milhares de fotografias, sons e documentos textuais baseados em imagem — não aparecem nos resultados das buscas feitas por tais mecanismos, e tampouco materiais e arquivos semelhantes de uma míriade de outras instituições. Os atuais mecanismos de buscas arranham a superfície do conteúdo da Rede (p. 107).

Esse ponto também foi suscitado por Han e Chang (2002):

> Em julho de 2000, os analistas estimavam que chegava a pelo menos 100 000 o número de bases de dados pesquisáveis na Rede. Essas bases de dados oferecem informações de alta qualidade, com boa manutenção, mas não são facilmente acessíveis. Como os atuais *crawlers* da Rede não podem consultar essas bases, os dados que contêm permanecem invisíveis para os mecanismos de busca tradicionais (p. 64).

Metadados

O termo 'metadados' possui várias definições possíveis. Cleveland e Cleveland (2001) lidam assim com essa questão:

> Repetidamente, definem-se metadados como *dados sobre dados*. Ainda que necessária, não é uma definição suficiente. Metadados quer dizer dados sobre dados que são estruturados para descrever um objeto ou recurso de informação. Caracterizam dados de fontes e descrevem suas relações. Autores de recursos, editoras, bibliotecários e outros profissionais da informação podem criar metadados. Podem estar incorporados ao recurso ou mantidos em repositórios separados de metadados (p. 223).

Hock (2001) prefere o termo 'metaetiquetas' que define como:

> A porção (campo) da codificação HTML de uma página da Rede que permite a quem a cria inserir texto que descreva o conteúdo da página. O conteúdo das metaetiquetas não aparece na página quando esta é visualizada na janela de um navegador (p. 220).

O chamado *Dublin Core* é um conjunto de itens de metadados (metaetiquetas) para descrever recursos disponíveis em rede. Tornou-se uma norma *de facto*

para descrição de recursos da Rede. Cleveland e Cleveland (2001) ressaltam que o Dublin Core:

> [...] proporciona informações de indexação para fontes documentais, inclusive indicadores para título, criador, assunto, descrição, editora, colaboradores, data, tipo, formato, identificador de recurso, língua, relação com recursos afins e gerenciamento de direitos autorais. [...]
>
> O conceito de *core* [núcleo] refere-se a um consenso alcançado por profissionais da informação e especialistas de assuntos sobre quais os elementos que são essenciais ou fundamentais para manter representações de informação, principalmente em formatos eletrônicos. [...]
>
> Uma das finalidades do desenvolvimento do Dublin Core era criar um esquema alternativo para as complexas técnicas de catalogação e fosse utilizável por catalogadores, não-catalogadores e especialistas em buscas de informação. Os criadores de bases de dados eletrônicas disporiam, em certo sentido, de catalogação do tipo 'faça você mesmo', mediante o preenchimento de espaços em branco. Os consulentes poderiam usá-la para navegar nas disciplinas e através delas, em ambiente internacional, na Rede (p. 224).

Hearst (1999) distingue entre 'metadados externos' e 'metadados de conteúdo'. Definem-se os primeiros como os dados 'relativos à produção e utilização do documento', como autor, lugar de publicação e data de publicação. 'Metadados de conteúdo', naturalmente, são os dados relativos ao conteúdo (assunto) do documento. É claro que este livro diz mais respeito aos metadados de conteúdo.

Outra distinção foi feita por O'Neill et al. (2001). Os tipos de metadados que estes autores reconhecem são: "aquilo que é explicitamente fornecido pelo autor do documento da Rede, e aquilo que é proporcionado automaticamente pelo editor de HTML com que o documento é criado". Com base numa amostra de registros da Rede, colhida em junho de 1998, concluem:

> Os resultados [...] sugerem que a utilização de metadados é bastante comum em documentos da Rede. No entanto, várias ressalvas devem ser feitas a essa conclusão. Primeiro, é evidente que grande parte do uso atual dos metadados pode ser atribuído à geração automática de metaetiquetas pelos editores de HTML. Não está claro que essa espécie de metadado seja particularmente útil para facilitar a descoberta e a descrição de recursos. Segundo, com freqüência os metadados são usados para descrever apenas o próprio sítio, ou, no máximo, um pequeno subconjunto dos documentos do sítio. Os atuais padrões de uso dos metadados estão muito distantes da descrição exaustiva do documento em nível de página. Finalmente, a maior parte da utilização dos metadados ainda é casuística; com poucas exceções, a maior parte dos sítios não obedece a um conjunto bem-definido de elementos de metadados (p. 374).

Constatou-se que cerca de 17% da amostra dos sítios continham 'palavras-chave'. Porém, não eram necessariamente termos muito úteis para recuperação:

> A característica mais notável foi que as palavras-chave, embora normalmente pertinentes, de algum modo, ao conteúdo do sítio, eram, não obstante, muitas vezes extremamente genéricas. Por exemplo, o sítio de uma universidade na Rede teria

> 'educação' como palavra-chave, ou um provedor da internet usaria 'Rede' como palavra-chave. A utilização de palavras-chave dessa maneira sugere que a finalidade dos metadados é aumentar ao máximo as possibilidades de a relevância do sítio ser percebida nas consultas feitas pelos mecanismos de buscas, ao invés de ajudar no descobrimento do sítio por si mesmo ou como membro de um conjunto relativamente pequeno de resultados de consultas de busca. Naturalmente, a generalidade de algumas dessas palavras-chave pode ser atenuada pela combinação de duas ou mais de duas numa consulta de busca. Além disso, não é necessariamente o caso de terem sido escolhidas palavras-chave não-específicas para aumentar a probabilidade de recuperação no sítio. É provável que, em alguns casos, o uso de termos extremamente gerais seja simplesmente o resultado de uma prática de indexação ruim (p. 366).

Dempsey e Heery (1998) chamaram atenção para a crescente importância dos metadados:

> Os metadados difundir-se-ão nos ambientes viáveis de informação digital a tal ponto que [...] será difícil falar genericamente sobre eles. As análises sensatas cingir-se-ão ao uso dos metadados para fins específicos ou em comunidades específicas (p. 168).

A importância dos metadados para arquivos digitais de vídeo foi analisada por Wactlar e Christel (2002).

Drott (2002) estudou a extensão com que os sítios de grandes empresas na Rede incluem 'recursos auxiliares de indexação' (isto é, auxílios no texto para orientar os robôs sobre o que procurar para fins de indexação). Ele examinou tanto os auxílios positivos (metaetiquetas incorporadas que identificam 'palavras-chave' ou 'descrição' no texto) e negativos (uso de um arquivo robots.txt que pode impedir que um robô indexe uma parte de um sítio da Rede). Entre 2000 e 2002 ele detectou um aumento no emprego de metaetiquetas.

Alguns autores têm chamado atenção para o fato de que os metadados tanto podem ter desvantagens quanto vantagens. DeRuiter (2002) é um deles:

> Para orientar mecanismos de busca sem confundir as pessoas, certas informações foram colocadas em metaetiquetas que não são imediatamente visíveis na apresentação de uma tela na Rede. Comprovou-se que isso era uma vantagem discutível. Por um lado, um mecanismo de busca pode encontrar a informação com eficiência, mas, por outro lado, muitas vezes não fica claro para os usuários por que uma página apareceu na busca (p. 205).

Craven (2001a) examinou a estabilidade dos metadados na Rede. Ele asssim resumiu os resultados que obteve:

> Quatro conjuntos de páginas da Rede anteriormente visitadas no verão de 2000 foram revisitadas um ano depois. De 707 páginas que, no ano de 2000, continham descrições de metaetiquetas, 586 permaneciam com essas descrições em 2001, e, de 1 230 páginas que careciam de descrições em 2000, 101 possuíam descrições em 2001. Nas páginas de abertura [*home pages*] parecia que tanto havia perdas quanto mudanças das descrições, mais do que nas outras páginas, com cerca de 19% das descrições modificadas nos dois conjuntos em que as páginas de abertura predominavam *versus* cerca de 12% nos outros dois conjuntos (p. 1).

Em estudo relacionado a esse (Craven, 2001b), ele examinou a aparência das 'descrições' (essencialmente um tipo de resumo) em metadados da Rede. É assim que ele descreve os resultados alcançados:

> Amostras aleatórias de 1 872 páginas da Rede registradas no Yahoo! e 1 638 páginas localizáveis a partir de páginas registradas no Yahoo! foram analisadas quanto ao uso de metaetiquetas, especialmente as que continham descrições. Setecentas e vinte e sete (38,8%) das páginas registradas no Yahoo! e 442 (27,0%) das outras páginas incluíam descrições em metaetiquetas. Algumas das descrições excediam grandemente as diretrizes usuais relativas à extensão de 150 ou 200 caracteres. Um número relativamente pequeno (10% das páginas registradas e 7% das demais) duplicavam exatamente a redação encontrada no texto visível; a maioria repetia algumas palavras e expressões. Ao contrário das orientações documentadas dadas aos redatores de páginas da Rede, era menos provável que as páginas com menos texto visível tivessem descrições. Era mais provável que as palavras-chave aparecessem mais perto do começo de uma descrição do que mais perto do fim. Eram mais comuns sintagmas nominais do que frases completas, especialmente em páginas não-registradas (p. 1).

Uma importante iniciativa para aplicação de metadados a recursos da Rede foi estabelecida com a denominação de CORC (Cooperative Online Resource Catalog), um programa conjunto do OCLC e um grande grupo de bibliotecas participantes. Em 2002 foi rebatizado para 'Connexion'. Compreende uma base de dados de descrições de recursos da Rede e uma base de dados de 'desbravadores' [*pathfinders*] que oferece acesso navegável aos recursos por meio da Classificação Decimal de Dewey (CDD) (Vizine-Goetz, 2001; Hickey e Vizine-Goetz, 2001). É possível atribuir automaticamente, por meio do programa Scorpion, números da CDD aos recursos selecionados pelas bibliotecas participantes para inclusão na base de dados (Shafer, 2001). O Scorpion funciona procurando as melhores coincidências entre expressões-chave no texto e o texto associado aos números da CDD. As atribuições feitas pelo Scorpion podem ser consideradas sugestões a serem submetidas à revisão de pessoas. Embora freqüentemente sejam 'corretas' (isto é, estejam de acordo com a classificação feita por pessoas), nem sempre serão. Goodby e Reighart (2001a) descrevem sua pesquisa sobre aplicação de indexação automática (o sistema WordSmith) a registros CORC. O WordSmith pode selecionar expressões candidatas em documentos da Rede para que sirvam como possíveis termos de indexação, apresentando-os como sugestões ao catalogador que estiver criando um registro CORC. Até outubro de 2002, umas 500 instituições haviam contribuído com cerca de 700 000 registros.

Outra iniciativa relacionada a filtros de qualidade na Rede é o Open Directory Project (<http://dmoz.org>), que pretende ser um cadastro de recursos da Rede, selecionados por serem de boa qualidade, numa ampla variedade de áreas temáticas. É mantido graças aos esforços de voluntários que se dedicam a selecionar sítios em suas áreas de conhecimento. Em 18 de janeiro de 2003 o projeto declarava incluir mais de 3,8 milhões de sítios selecionados por mais de 54 000 colaboradores e organizado em mais de 460 000 categorias.

Naturalmente, há muito que a profissão de bibliotecário lida com metadados — na forma de entradas descritivas em catálogos em fichas, impressos e em linha. Não obstante, os metadados exigidos por recursos da Rede são, de algum modo, diferentes dos metadados tradicionalmente utilizados para descrever livros e outros materiais impressos, inclusive porque têm de descrever coleções inteiras de registros ao invés de itens individuais (Hill et al., 1999) e que podem referir-se a objetos (por exemplo, peças de museus) ao contrário de texto (ver, por exemplo, Zeng, 1999). Ademais, talvez sejam necessários diferentes níveis de metadados para os mesmos materiais, a fim de atender às necessidades de diferentes públicos que podem até incluir crianças (ver, por exemplo, Sutton, 1999).

Encontra-se em Guenther e McCallum (2003) uma análise de desenvolvimentos recentes sobre os metadados, inclusive o MODS (Metadata Object and Description Schema) e METS (Metadata Encoding and Transmission Standard).

Resumos na Rede

Os metadados incluídos num sítio da Rede podem conter texto que é um tanto aparentado com um resumo — pelo menos uma anotação ou nota de conteúdo. Se não houver nada dessa espécie, alguns mecanismos de busca usarão as primeiras linhas do próprio texto como uma espécie de resumo.

Alguns dos serviços de busca constroem primeiro 'resumos' para os recursos que encontram e, em seguida, tornam pesquisáveis as palavras do resumo ou extraem palavras do resumo e não do texto completo. No entanto, as empresas nisso envolvidas costumam não ser muito informativas sobre como realmente funcionam seus processos automáticos de elaboração de resumos.

Wheatley e Armstrong (1997) salientaram que os recursos acessíveis na Rede podem exigir uma abordagem algo diferente da elaboração de resumos, em especial porque provavelmente se refiram a coleções de textos (ou, com efeito, imagens) e não itens individuais:

> Um 'resumo da internet' ideal incluiria, por exemplo, orientação ao usuário, avaliação da autoridade, análise de atributos físicos (o *design* do sítio ou a facilidade de navegação), juízos de qualidade, ou apontadores para fontes alternativas (p. 212).

Eles comparam resumos ou extratos extraídos da internet com resumos de bases de dados bibliográficos convencionais e resumos ou descrições de *gateways* da internet. Foram feitas comparações sobre legibilidade, conteúdo e estilo.

Alguns itens da Rede incluem palavras-chave ou expressões atribuídas pelos autores, as quais podem funcionar como sucedâneos para o objetivo de fazer buscas. No entanto, a grande maioria não o faz, o que estimulou a realização de pesquisas visando a extrair automaticamente tais expressões de textos da Rede, conforme foi visto no capítulo anterior (ver, por exemplo, Jones e Paynter, 2002).

Para uma análise das características de 'descrições' (um tipo de resumo) de páginas da Rede, ver Craven (2001b). Ele verificou, por exemplo, que muitas tendem a empregar sintagmas nominais ao invés de frases completas.

Spamming de índice e outras trapaças

Um problema potencialmente grave na Rede é causado pelo fato de que os desenvolvedores de sítios querem que eles sejam encontrados, o que é ainda mais verdadeiro quando há interesse de lucro. Alguns anos antes de a internet surgir, Price (1983) advertia que o acesso em rede a recursos eletrônicos poderia tentar os autores, pessoais e coletivos, a tornar seus trabalhos mais atraentes para os leitores ou mais recuperáveis. Isso é o que ocorre atualmente na internet num fenômeno que tem sido chamado de *spoofing*, ou *spamming* de índice.

Lynch (2001) tratou, com detalhes, do assunto da confiança e da procedência na internet. Ele observa que:

> Os documentos digitais em ambiente distribuído talvez não se comportem de modo coerente; como são mostrados tanto para pessoas que desejam vê-los quanto para sistemas de *software* que desejam indexá-los por meio de programas de computador, eles podem ser alterados, talvez de forma radical, para cada apresentação, que pode ser moldada para um receptor específico. Ademais, a informação que uma pessoa retira da apresentação de um documento por meio de programas, como um navegador da Rede, pode ser muito diferente da que um programa de indexação extrai até do mesmo documento-fonte, a não ser que o programa de indexação seja projetado para levar em conta o impacto do documento na percepção dos seres humanos. [...]
>
> Os sítios interessados em manipular os resultados do processo de indexação começaram rapidamente a explorar a diferença entre o documento como é visto pelo usuário e o documento como é analisado pelo *crawler* de indexação por meio de um conjunto de técnicas genericamente denominadas '*spamming* de índice'. Por exemplo, um documento pode ser abarrotado com milhares de palavras que o usuário não veria, porque se confundem com o fundo da página, numa fonte de tipos diminutos, porém que seriam encontradas pelo *crawler* de indexação. O resultado disso tem sido uma corrida armamentista entre indexadores e desenvolvedores de sítios da Rede, com os serviços de indexação acrescentando maior complexidade à extração de palavras, análise estatística, processamento da linguagem natural e outras tecnologias. Os serviços de indexação também complementam a indexação direta do conteúdo com informações contextuais, como, por exemplo, quantos outros sítios se vinculam com uma página, como uma forma de tentar identificar páginas importantes.
>
> É importante compreender que, quando um *crawler* requisita uma página para indexar, não se trata simplesmente de ler um arquivo em alguma espécie de sistema de arquivo de rede; ele está requisitando uma página para um servidor da Rede por meio do protocolo http. A requisição inclui identificação da fonte do pedido (em vários níveis — o programa que está pedindo e a máquina para a qual a requisição é enviada), e o servidor da Rede pode ser programado para responder de modo diferente a solicitações idênticas de diferentes origens. Os motivos para isso podem ser bastante generosos; por exemplo, alguns servidores oferecem páginas que são ajustadas, para indexar eficientemente, com os algoritmos de indexação usados por diferentes *crawlers*. Outros motivos para reações sensíveis à fonte são mais ativamente maldosos, como a prática do seqüestro de página [*pagejacking*]. Um exemplo tornará mais fácil a visualização do que se trata. Suponhamos que se tem um produto X que concorre com o produto Y fabricado por outra empresa. Quando as pessoas colocassem uma consulta nos mecanismos de busca da Rede perguntando por Y, sua vontade seria que o mecanismo de busca, ao contrário, respondesse com o envio de sua página com o anúncio de X. Você leva uma cópia da página para Y e fornece isso ao serviço de indexação da Rede, mas quando um usuário (ao contrário do serviço de indexação, clica no URL, você envia a página de seu produto X ao invés da página copiada para Y. A concorrência não é o único motivo; por exemplo, talvez você quisesse garantir que as páginas de uma organização de que você não gosta fossem devolvidas em resposta a pedidos de material de sexo explícito. O seqüestro de páginas é definido, geralmente, como o fornecimento arbitrário de documentos com entradas de índice arbitrárias e independentes. É claro que constitui um problema enorme a construção de sistemas de recuperação de informação capazes de fazer face a esse ambiente, e os *crawlers* da Rede estão começando a integrar uma grande variedade de controles de validade (como examinar redes de vínculos entre páginas e sítios) na tentativa de identificar e filtrar tentativas prováveis de seqüestro de páginas (p. 13-14).

Drott (2002) também se debruçou sobre o problema do *spamming*:

> Importante companhia de seguros imaginou a artimanha de repetir várias vezes a mesma palavra. Isto é, sua entrada de 'palavras-chave' repetia cada palavra seis vezes (por exemplo, prêmios prêmios prêmios prêmios prêmios prêmios). Esta espécie de tentativa de garantir melhor posição na indexação é com freqüência encontrada no texto de sítios pornográficos onde as palavras são ocultadas com o estratagema de usar uma mesma cor no texto e no fundo da página. Ressalva seja feita, a seu favor, que a companhia de seguros não implantou esse questionável dispositivo. A maioria dos serviços de busca desconta as repetições de modo que a repetição de palavras-chave é inútil. E, indo ao que interessa, é difícil imaginar uma razão comercial legítima para uma companhia de reputação querer envolver-se em tal prática (p. 214-215).

Introna e Nissembaum (2000) asseguram que, em termos de acesso, a Rede favorece os ricos, os inescrupulosos e os tecnologicamente proficientes. Os dois últimos podem promover o acesso a sítios da Rede por meio de *spamming* e outras vigarices. Os ricos podem pagar aos mecanismos de buscas para alcançar 'proeminência' no *ranking* dos resultados.

Mowshowitz e Kawaguchi (2002) alertaram para a existência de viés na Rede. Isto é, os mecanismos de busca podem criar viés na seleção de sítios a que proporcionam acesso. Exemplificam esse viés com buscas sobre produtos (uma busca sobre refrigeradores mostra viés para certos fabricantes em alguns mecanismos de buscas) e sobre eutanásia. Embora não procurem explicar por que ocorre o viés, a ele se referem como "um problema socialmente importante".

Floridi (1996) talvez haja sido o primeiro a apontar os perigos da internet como fonte de desinformação. Desde então o tema foi tratado com detalhes em livro organizado por Mintz (2002), que a ele se refere como malinformação [*misinformation*], que é definida como informação "intencionalmente errada ou equivocada". Exemplos incluem informações médicas e comerciais que carecem de credibilidade, *scams* para arrecadação de dinheiro para fins pseudocaritativos, trapaças por correio eletrônico, e orientação jurídica perigosa.

Embora os problemas abordados nesta seção não sejam, de per si, questões

de indexação, ilustram com clareza a necessidade de filtros de qualidade e o fato de um pedaço substancial da Rede não merecer o mínimo de atenção em matéria de indexação.

Vinculação de hipertexto/hipermídia

É óbvio que, na estrutura da Rede, acha-se implícita uma forma de indexação. O fato de duas fontes, *A* e *B*, estarem vinculadas na Rede implica que ambas pertencem, em algum sentido, à mesma classe, e que os termos relativos a *A* também podem ser úteis na recuperação de *B*, e vice-versa (Savoy, 1995).

É natural, portanto, que os problemas de indexação relativos a fontes em hipertexto e hipermídia hajam recebido grande atenção. Alguns trabalhos (por exemplo, Salton e Buckley, 1992; Savoy, 1995; Salton et al., 1997) analisam processos para estabelecimento automático de vínculos de hipertexto. Isso também é tema de livro organizado por Agosti e Smeaton (1996). Agosti et al. (1995) descrevem métodos para estabelecimento automático de vínculos de hipermídia, com o uso de critérios de associação estatística, em tempo real, pesquisando interação com a Rede. Referem-se a isso como 'autoria automática' de hipermídia. Embora não proponham explicitamente isso, os vínculos de hipertexto estabelecidos pelo consulente, se forem registrados, serão úteis para consulentes posteriores. A idéia é conceitualmente similar à de um 'tesauro em crescimento'.

Arents e Bogaerts (1996) fazem revisão da literatura sobre recuperação de hipermídia. Embora mencionem amiúde a 'indexação', muitos dos métodos (na maioria experimentais) que estudam envolvem pesquisas ou 'navegação' em hiperredes, seguindo vínculos preestabelecidos ou formados no próprio processo de busca. Os 'navegadores' gráficos ou 'mapas' destinados a fornecer ao usuário um panorama visual dos vínculos na rede (ver Zizi, 1996, por exemplo) lembram os 'mapas semânticos' propostos por Doyle (1961) há mais de 40 anos.

Como foi indicado neste livro, o hipertexto e as redes de hipermídia criam fontes de informação que não possuem fronteiras claramente definidas. Chiaramella e Kheirbek (1996) tratam dessa questão. Salientam que "os documentos não são mais unidades atômicas" o que faz com que mudem nossas concepções sobre o que constitui não apenas 'documento' mas '*corpus*' e também 'índice'.

Tessier (1992), Ellis et al. (1994, 1996) e Chu (1997) são alguns autores que abordaram as relações indexação/hipertexto. Tessier examinou as semelhanças entre vinculação de hipertexto e indexação convencional. Ela sustenta que os autores de hipertexto vinculam o texto de forma muito parecida com a forma como seriam vinculados na indexação convencional. Ellis et al. constataram que seres humanos, solicitados a inserir vínculos de hipertexto numa coleção de textos, como indexadores convencionais, não revelam grande coerência nessa tarefa. Em artigo posterior (Ellis et al., 1996), testam o efeito dessa coerência da vinculação na eficácia da recuperação. Chu (1997) tentou aplicar os princípios da exaustividade e especificidade a vínculos de hipertexto. Embora a medida do primeiro seja precisa (número de vínculos por número de palavras num documento), a medida de especificidade é muito mais difícil de aplicar com êxito.

Srinavasan et al. (1996) focalizam os problemas da indexação e recuperação de itens na Rede. Seu trabalho sugere que técnicas que funcionam bem em ambientes mais estáveis (por exemplo, freqüência inversa de termos para a classificação dos itens recuperados) podem ser menos eficazes em "contexto [tão] heterogêneo e dinâmico". Alguns problemas relativos ao uso de uma base de dados de hipertexto em atividades de recuperação de informação são examinados por Dimitroff e Wolfram (1993). Mais tarde, Wolfram (1996), usando três modelos diferentes, pesquisou os vínculos entre registros de hipertexto.

Blair e Kimbaugh (2002) exaltam as virtudes de 'documentos exemplares' no projeto de sistemas de recuperação. Documentos exemplares são aqueles (como artigos de revisão, manuais clássicos e normas jurídicas) que melhor "descrevem ou exibem a estrutura intelectual de determinado campo". Sugerem vários usos possíveis (um é uma fonte de terminologia representativa para o campo), inclusive sua utilidade possível para o usuário da Rede. Se quem faz a busca recupera primeiro um documento exemplar, os vínculos de hipertexto nele inseridos podem ser usados para estender uma busca em várias direções.

Melucci (1999) avaliou a eficácia da recuperação de vínculos de hipertexto construídos automaticamente, e Blustein e Staveley (2001) oferecem uma revisão de trabalhos sobre geração e avaliação de hipertexto.

Um tanto relacionada com o hipertexto está a possibilidade para usuários dos recursos da Rede de nele fazer anotações, exatamente da mesma maneira que fariam anotações nas páginas de um livro didático. Essa possibilidade é examinada mais detidamente por Marshall (2000).

Classificação na internet

Vários mecanismos de buscas incluem alguma modalidade de categorização dos recursos a que proporcionam acesso. Esses 'cadastros' ['*directories*'] empregam alguma forma de classificação hierárquica. Em alguns casos, vários mecanismos de buscas compartilham o uso de um cadastro produzido alhures. O Yahoo! considera-se basicamente um cadastro da Rede, embora seja, essencialmente, uma combinação de cadastro/mecanismo de busca.

Além disso, algumas instituições vêm trabalhando com vista à organização de recursos da internet por meio de alguma modalidade de classificação. O OCLC, por exemplo, conta com várias iniciativas relacionadas com isso. Uma delas, Scorpion, é um sistema experimental para atribuição automática de números da Classificação Decimal de Dewey a recursos da Rede (Vizine-Goetz, 1998, 2001; Hickey e Vizine-Goetz, 2001). A base de dados NetFirst do OCLC (já desativada) também usava a estrutura hierárquica da classificação de Dewey, a fim de proporcionar acesso a recursos selecionados da Rede.

Zins (2002) examinou os tipos de classificação usados em importantes por-

tais e cadastros classificados da Rede. Identificou oito princípios de classificação adotados nesses recursos, e considerou cinco deles 'relativos a conteúdo': assuntos, objetos (por exemplo, pessoas e organizações), aplicações (por exemplo, compras), usuários (para quem se destina um recurso) e localizações (lugar), e os outros três princípios relativos a formato: mídia (por exemplo, imagens), 'referência' (por exemplo, dicionários, mapas), e línguas. Zins sugere a necessidade de integrar esses princípios numa classificação facetada para aplicação na internet. Embora tal integração seja teoricamente atraente, Zins parece subestimar o fato de que a classificação dos recursos da Rede é essencialmente pragmática e prática, e que os esquemas unidimensionais hoje empregados talvez sejam tudo de que os usuários precisam para utilização bem-sucedida da Rede.

Algumas bibliotecas começam a atribuir números de classificação aos recursos da Rede a que proporcionam acesso. Elrod (2000) resumiu um debate em linha sobre essa questão. Um dos colaboradores apresentou uma justificativa boa e precisa dessa prática:

> Ao atribuir números de classificação a materiais acessíveis por intermédio de seu catálogo em linha, mas que não se encontram fisicamente alojados em suas estantes (recursos da internet), e seu catálogo em linha de acesso público permite pesquisa por número de chamada, então o cliente pode 'pesquisar' não apenas o material que você tem em seu acervo, mas também aqueles recursos da internet sobre o mesmo assunto ou assuntos próximos. Uma vez que o mesmo cliente pode acessar o recurso da internet por intermédio do vínculo fornecido no registro bibliográfico, ter esse registro destacado numa busca de números de chamada constitui método pelo qual podemos oferecer maior acesso à informação (Elrod, p. 23).

Enquanto os mecanismos de busca da Rede proporcionam acesso em nível de página, os cadastros e portais geralmente fornecem acesso em nível de sítio da Rede. Casey (1999) examinou a necessidade de um índice analítico da Rede, ou seja, que empregue alguma forma de classificação ou outro vocabulário controlado para indexar recursos abaixo do nível de sítio. Ela reconhece "a impossibilidade de um índice analítico exaustivo da internet", mas acredita que a "criação de pequenos índices focais pode ser a melhor solução para acessar tipos específicos de informação digital". Isso é precisamente o que os portais examinados na seção seguinte objetivam fazer.

Muitos autores enfatizaram a necessidade de mais classificação dos recursos da Rede. Trippe (2001) assim se expressou:

> Segundo alguns, o caminho que leva a uma melhoria da recuperação da informação na Rede está em taxonomias aplicadas de forma inteligente. Segundo tal opinião, é preciso identificar com mais precisão o conteúdo, mediante o uso de categorias, de modo que os mecanismos de busca e outros auxílios à navegação possam ser mais bem sintonizados para ajudar o usuário. Como, cada vez mais, os conteúdos caminham rumo à Rede, essas fontes de dados precisam beneficiar-se das tecnologias e técnicas que permitem às pessoas visualizar, navegar e procurar dados por meio de categorias que sejam amplamente compreendidas (p. 44).

E passa a descrever vários produtos comerciais que se destinam a realizar automaticamente várias atividades de categorização em fontes da Rede.

Portais

Embora bases de dados bibliográficos, como as da National Library of Medicine, estejam acessíveis na internet, a grande maioria dos recursos da Rede não está 'indexada' no sentido com que a palavra é empregada neste livro, isto é, pela atribuição de termos, feita por seres humanos ou computador, talvez extraídos de um vocabulário controlado. Não obstante, bibliotecas especializadas e centros de informação podem oferecer um serviço importante com a identificação dos recursos da Rede de maior relevância e utilidade para seus usuários, indexando de alguma forma esses recursos, e desenvolvendo um *gateway* que proporcione acesso a eles por meio dos elementos de metadados. Vários desses *gateways* ou 'portais' são descritos e exemplificados em Wells et al. (1999), que a eles se referem como 'bibliotecas virtuais'.

Um *gateway* ou portal típico dessa categoria é o EEVL, que constitui um empreendimento conjunto de várias universidades do Reino Unido. Segundo Breaks e Guyon (1999), trata-se de:

> [...] um *gateway* para sítios da internet, de qualidade, sobre engenharia [...] [que] tem por objetivo permitir que professores, pesquisadores e estudantes de engenharia no Reino Unido utilizem melhor os recursos disponíveis na internet graças à melhoria do acesso a tais recursos. Alcançamos isso por um processo de identificação, filtragem, descrição, classificação e indexação de sítios de qualidade antes que sejam acrescentados a uma base de dados livremente disponível na Rede Mundial (p. 76).

A base de dados contém descrições pesquisáveis e vínculos com sítios da internet que tenham interesse. Os recursos são categorizados com um esquema de classificação especialmente desenvolvido para tal fim. A sigla EEVL (<http://www.eevl.ac.uk/>) originalmente significava Edinburgh Engineering Virtual Library. Posteriormente foi renomeada Enhanced and Evaluated Virtual Library quando seu campo de ação foi ampliado para incluir matemática e informática. Em 21/9/2002, a EEVL proporcionava acesso a mais de 9 000 sítios.

Um portal parecido é o Agriculture Network Information Center (AGNIC), mantido pela National Agricultural Library e várias outras instituições. O AGNIC (<http://www.agnic.org/>) proporciona acesso a recursos da Rede em 15 categorias gerais, todas relativas à agricultura em seu sentido mais amplo.

INFOMINE (<http://infomine.ucr.edu/>) descreve-se assim:

> [...] uma biblioteca virtual de recursos da internet, relevantes para os corpos docente e discente e pesquisadores da universidade. Contém recursos úteis da internet, tais como bases de dados, periódicos eletrônicos, livros eletrônicos, quadros de avisos, listas de endereços, catálogos de bibliotecas [...] em linha, artigos, cadastros de pesquisadores e muitos outros tipos de informação.
>
> O INFOMINE foi feito por bibliotecários. Profissionais de universidades e facul-

dades, como University of California, Wake Forest University, California State University, The University of Detroit - Mercy, colaboraram na construção do INFOMINE.

Em 21/9/2002, anunciava que proporcionava acesso a mais de 23 000 recursos. Os sítios incluídos são comentados e recebem cabeçalhos de assuntos (Library of Congress) para melhorar o acesso (Mitchell e Mooney, 1999).

Outros portais destinam-se a facilitar o acesso a recursos da Rede que sejam de interesse potencial para usuários de bibliotecas públicas. O Librarians' Index to the Internet (<http://lii.org/>) é assim descrito por Hinman e Leita, 1999:

> um cadastro temático, pesquisável e comentado, de [...] recursos da internet, selecionados e avaliados quanto à sua utilidade para as necessidades de informação do usuário de bibliotecas públicas. Os recursos são selecionados e indexados por uma equipe de bibliotecários treinados, voluntários, de bibliotecas da Califórnia (p. 144).

São abrangidos mais de 10 000 recursos da internet, organizados em categorias e subcategorias. São empregados cabeçalhos de assuntos da Library of Congress, com modificações.

O Getty Information Institute é outra instituição atuante nesse tipo de esforço. Busch (1998) descreveu como os vocabulários controlados do Getty podem ser usados para proporcionar melhor acesso a recursos em arte.

Portais desse tipo são importantes como filtros dos recursos em rede. O componente de 'valor agregado' ou 'filtro de qualidade' — de seleção, anotação, indexação — é de suma importância. Esse ponto é enfatizado na EEVL:

> As buscas na EEVL recuperarão recursos de alta qualidade, mas, em virtude de os recursos constantes da EEVL serem colhidos à mão, aqui você não encontrará tantos recursos quantos encontraria em alguns serviços, porém serão os melhores!

Os portais mencionados neste capítulos destinam-se a serem acessados por uma grande variedade de usuários potenciais. No entanto, são possíveis portais mais restritos e especializados. As bibliotecas podem criar seus próprios portais para recursos da Rede. Hurt e Potter (2001) dão um exemplo:

> No campus da Georgia State University, os bibliotecários de ligação (que também são bibliotecários de referência e desenvolvimento de coleções) dedicam-se ativamente à identificação e criação de sítios na Rede, particularmente em suas especialidades, e desenvolvem sítios sobre vários temas, que incorporam outros sítios bem-conceituados. Outro grupo importante de bibliotecários muito envolvidos com a Rede são docentes bibliotecários de coleções especiais e arquivos, muitos dos quais criam arquivos digitais para ampliar o conteúdo da biblioteca virtual na Rede (p. 23).

Medeiros et al. (2001) descrevem a abordagem de uma biblioteca universitária de medicina em relação a um portal, que utiliza como base o Cooperative Online Resource Catalog (CORC). Assim se referem às vantagens disso:

> A biblioteca pode usar o CORC para selecionar sítios que ofereçam conteúdo de qualidade. O cliente da biblioteca é atendido ao poder dirigir-se, sem esforço, ao recurso exato de que precisa e evitar ficar vasculhando os resultados dos mecanismos de buscas que freqüentemente consistem em páginas de vínculos irrelevantes. As características típicas do CORC, como controle de autoridade para acesso a nomes, ajudam na localização dos recursos (p. 112).

Um centro de informação industrial pode desenvolver um portal que aponte para recursos de maior interesse e utilidade para a empresa e integrar isso com a própria intranet da empresa, adotando o mesmo modo de acesso temático (por exemplo, esquema de classificação). Ver, por exemplo, Crandall (2000). Bannan (2002) trata do tema dos portais de empresas, mas, em sua opinião, eles proporcionam acesso a informações internas, e possivelmente permitem a pessoas de fora o acesso a recursos selecionados da empresa, ao invés de *gateways* para informações úteis (para a empresa) alhures na Rede.

Campbell (2000) descreveu sua visão de um 'portal de cientistas' destinado a:

> promover o desenvolvimento e proporcionar acesso a conteúdos da mais alta qualidade na Rede. Facilitaria o acréscimo de material de alta qualidade ao promover padrões, fazer buscas em bases de dados, e oferecer uma variedade de ferramentas de apoio. Com isso, bibliotecas, empresas e muitas outras organizações estariam capacitadas a contribuir para uma biblioteca digital acessível e distribuída (p. 3)

Embora Campbell não trate diretamente das questões relativas à indexação, elas estão implícitas no reconhecimento de que o portal "também ofereceria excelentes tesauros eletrônicos que orientariam, com precisão, os pesquisadores para áreas de interesse". A Association of Research Libraries vem atuando no desenvolvimento dessa idéia por intermédio do Scholar's Portal Project (<http://www.arl.org/access/scholarsportal/>). Ver Jackson (2002) para os avanços nessa área até meados de 2002.

Em Awre e Wise (2002) encontra-se uma breve revisão de desenvolvimentos recentes relativos a portais no Reino Unido.

Place (1999) faz uma previsão do futuro dos *gateways* temáticos:

> Os usuários já podem aproveitar os *gateways* temáticos, que, juntos, descrevem dezenas de milhares de recursos de alta qualidade na internet. No futuro, os usuários verão os atuais *gateways* temáticos crescer notavelmente de tamanho, à medida que mais bibliotecários e profissionais da informação contribuírem para eles e à medida que soluções automatizadas e humanas para descobrimento de recursos forem integradas. Verão também o surgimento de novos *gateways* e poderão fazer buscas cruzadas simultanea e inconsutilmente em diferentes *gateways*.
>
> Também no futuro, será possível usar perfis de usuários para habilitar os *gateways* temáticos a fornecer um serviço de informação personalizado. Os usuários serão solicitados a inserir numa base de dados suas preferências em matéria de informação, de modo que os *gateways* possam notificá-los sobre novos recursos que surjam no catálogo (p. 243-244).

O futuro da indexação e da redação de resumos é examinado, mais detidamente, no capítulo seguinte, e último, deste livro.

Capítulo 17

O futuro da indexação e redação de resumos

Escrevendo há quase 50 anos, Fairthorne (1958) afirmou que "A indexação é o problema fundamental bem como o obstáculo mais dispendioso da recuperação da informação". E a indexação continua sendo o problema principal do acesso à informação, e a mente de Fairthorne por certo teria ficado atônita diante da imensidão dos problemas de acesso à informação suscitados pela Rede Mundial.

Missingham (1996) oferece uma clara explanação desses problemas:

> Não se pode considerar a internet como se fosse apenas mais um passo na história da indexação. Ela suscita enormes desafios e exige uma abordagem muito diferente da indexação para alcançar uma recuperação eficiente da informação.[...] A indexação da internet oferece muitos desafios: ela contém milhões de documentos ou arquivos; a localização desses documentos/arquivos muda freqüentemente; não há nenhum controle de qualidade da informação na internet, nenhuma coerência no uso da terminologia, ou mesmo no emprego de títulos; é muito difícil manter-se a par das novas fontes; os índices são complicados porque muitos dependem de informações comunicadas pelos próprios editores (algo parecido com o atual processo de catalogação na publicação). [...] Não há normas que exijam que sejam usados os autores ou os títulos, nem a exigência de que a informação principal inclua o título ou o subtítulo. A indexação da internet é, portanto, muito diferente da indexação de um artigo de periódico, onde essas informações identificadoras normalmente são claras (p. 35).

E ela acrescenta que o maior de todos os problemas talvez seja:

> [...] a natureza volátil da rede onde indexar um recurso é realmente como se se estivesse a enxugar gelo, pois hoje ela pode estar ali e amanhã já ter desaparecido ou mudado completamente. Não só o nome, o conteúdo e a localização do recurso podem alterar-se regularmente, mas também sua acessibilidade e formato mudar facilmente (p. 36).

O grande defeito da internet como fonte de informação, fora seu tamanho, está no fato de que ela carece de qualquer forma de controle de qualidade. O fato de os serviços de informação funcionar com razoável eficiência no mundo do papel impresso deve-se a que várias instituições existem para desempenhar a função de filtro de qualidade. As editoras de livros e periódicos científicos adotam processos de revisão/avaliação que são, pelo menos em certa medida, eficazes para eliminar a maior parte do que é imprestável. Os serviços que editam índices e resumos proporcionam o nível seguinte de filtro de qualidade, principalmente ao escolher os periódicos, séries de relatórios ou outras publicações que serão analisados regularmente. Por fim, as bibliotecas, particularmente as que servem às comunidades de ensino e pesquisa, colocam os filtros mais perto dos usuários reais quando compram materiais considerados de maior utilidade para esses usuários e quando organizam as coleções segundo níveis de acessibilidade, para que fiquem mais próximos (fisicamente e talvez também intelectualmente) os materiais que mais provavelmente os usuários venham a precisar.

É claro que a imensa vastidão de recursos mal-organizados que estão acessíveis, pelo menos em sentido teórico, na internet, faz com que a construção de filtros eficazes seja uma proposta intimidadora, tanto para pessoas quanto para instituições. Ademais, dão-nos a certeza de que a situação haverá de ficar muito pior (Weld et al., 1995).

Embora muitos documentos da Rede sejam de baixa utilidade, outros podem simplesmente desaparecer, conforme salientou Missingham. Spinellis (2003) constatou que cerca de 28% dos URLs referenciados em dois importantes periódicos de ciência da computação, entre 1995 e 1999, não estavam mais acessíveis em 2000, número esse que aumentou para 41% em 2002. O índice de desaparecimento de documentos da Rede pode, *grosso modo*, equivaler ao índice de obsolescência da literatura de ciência da computação (isto é, declínio de uso com a idade). Não obstante, o fato de itens desaparecerem ou talvez reaparecerem em outro formato sem referência ao original não estimula investimento numa indexação dispendiosa.

Hoje em dia, não parece provável que a situação caótica causada pelo fenômeno do 'cada um será seu próprio editor' seja reversível. Em outras palavras, é difícil visualizar a possibilidade de que alguém poderia impor ou imporia normas de qualidade total à publicação ou distribuição através de redes. Por conseguinte, a viabilidade de uma vasta rede como recurso de informação dependerá da imposição de filtros de qualidade similares aos do mundo da impressão em papel.

Não há dúvida que a função de filtro é tão importante no ambiente eletrônico quanto o era num ambiente editorial dominado pela impressão em papel. Como indexação e resumos, numa ou noutra forma, são elementos essenciais na filtragem da informação, conclui-se que terão futuro. As perguntas que permanecem sem resposta, então, são as seguintes:

1. Qual a forma que terão essas atividades, e
2. A quem caberá ou caberia realizá-las?

É interessante observar que Odlyzko, que há alguns anos previu que tanto as bibliotecas quanto os periódicos científicos tornar-se-iam obsoletos, pelo menos em seu formato tradicional (ver, por exemplo, Odlyzko, 1995), é bastante positivo no que tange ao futuro dos serviços de indexação e resumos. Ele afirma (Odlyzko, 1999) que esses serviços sobreviverão porque sua contribuição intelectual é substancial e porque, por isso, são comparativamente baratos.

Jacsó (2002) discorda um pouco em relação aos serviços, mas continua sendo um firme adepto da necessidade de resumos na Rede:

> A crescente disponibilidade de bases de dados de texto completo fez diminuir a importância de bases de dados de resumos e índices nos últimos 10 a 15 anos, mas não a necessidade de resumos. As bases de texto completo precisam de resumos para que seu uso seja eficiente. A razão óbvia disso está em que passar os olhos nas listas com os resultados de buscas que contenham breves resumos ajuda tremendamente a selecionar os documentos-fonte mais promissores, mesmo quando os resumos deixam muito a desejar.

E acrescenta:

> A razão menos óbvia para a existência de resumos nessas bases de dados está em que se a busca se limitar ao campo do resumo numa base de dados de textos completos haverá garantia de que ela será mais precisa do que se for feita em centenas de milhares de documentos de texto completo (p. 22).

É claro que Jacsó não está se referindo necessariamente a resumos preparados por seres humanos, mas a resumos ou extratos preparados automaticamente. De fato, seu artigo passa em revista programas disponíveis comercialmente destinados à 'sumarização de documentos'.

Mani (2001) é outro autor que acentuou a importância da sumarização:

> A explosão da Rede Mundial trouxe consigo um estoque imenso de informações, em sua maior parte relativamente não-estruturadas. Isso fez surgir a demanda por novas maneiras de gerenciar esse corpo bastante sobrecarregado de informações dinamicamente cambiantes. Em tal ambiente, parece indispensável alguma forma de sumarização automática. Usuários da Rede, fontes de informação em linha e novos dispositivos móveis, além da necessidade da gestão do conhecimento pelas empresas, vêm exercendo pressão crescente em prol de avanços da tecnologia na questão da sumarização. Empresas comerciais passam cada vez mais a oferecer recursos de sumarização de textos, muitas vezes integrados com ferramentas de recuperação da informação (p. 529).

As propostas referentes à indexação de recursos da Rede abrangem um leque extremamente variado, inclusive afirmativas de que isso não é possível de modo algum. Por exemplo, Wellisch (1994) sustentou que "É improvável que os periódicos eletrônicos sejam indexados, devido à instabilidade de seus textos". Tendo em vista que a maioria das fontes na internet são muito menos estáveis do que os periódicos, ele provavelmente acha que todo o esforço — ou seja, a indexação de textos que estão sujeitos a freqüentes mudanças — seja uma causa perdida.

É evidente que a indexação profissional, feita por seres humanos, de toda a Rede é totalmente inviável. Mesmo que o fosse, grande parte do que aparece na Rede é de valor muito passageiro ou de qualidade excessivamente baixa para chegar a merecer tais cuidados com a indexação. Já uma indexação seletiva, profissional, é, naturalmente, viável. Owen (1994) e Weinberg (1996) são dois

autores que defenderam a indexação profissional numa base seletiva. Weinberg recomendou especificamente uma indexação no estilo dos índices de final de livro, e essa espécie de indexação poderia certamente ser aplicada a sítios específicos da Rede. Na realidade, já foi aplicada dessa forma, e Browne (2001) analisou e exemplificou os respectivos processos. Casey (1999) admite que o sonho que ela alimentava de um 'índice analítico' completo da Rede (isto é, que indexasse abaixo do nível de sítio) era utópico e que "índices pequenos, focais, talvez sejam a melhor solução".

Ellis et al. (1998) sugerem que um grande problema de qualquer abordagem relacionada com a indexação da Rede é o fato de que o indexador sempre estará muito longe do usuário:

> [...] na Rede Mundial [...] não há proximidade alguma entre projetista ou criador (que poderia ser qualquer um) e usuário potencial (que poderia ser qualquer um ou todos). Isso é agravado pela falta de uma noção clara por parte da maioria dos consulentes sobre o que é que os diversos mecanismos de buscas na realidade fazem quando realizam uma busca. De modo que a origem real dos problemas que ocorrem nas buscas feitas em fontes distribuídas em linha ou na internet não está nos problemas técnicos de indexação, mas na facilidade de acesso proporcionado por serviços em linha e a Rede Mundial a informações selecionadas, estruturadas e indexadas para um grupo de usuários (que possuem um conjunto de características e necessidades de informação) por espécies de usuários totalmente diferentes com características e necessidades totalmente diferentes.
>
> É natural que isso venha a exacerbar problemas existentes em relação à coincidência de conceitos entre indexador e usuário, pois os usuários encontram muitos arquivos ou sítios diferentes, com características, práticas de indexação e vocabulários diferentes, nenhum dos quais, com certeza, poderá satisfazer a todas ou mesmo algumas das necessidades de um usuário ou grupo de usuários potenciais. Essa é uma questão importante, pois os usuários mais distantes são, no que concerne a características e necessidades de informação, dentre os tipos de usuários imaginados e levados em conta pelos que criam ou indexam uma base de dados, os que mais provavelmente terão problemas em acessar informações relevantes dessa base de dados. O problema é a indexação para o usuário desconhecido (p. 44).

Pode-se considerar os documentos da Rede como 'dinâmicos', e não estáticos, no sentido de que podem ser modificados pelo seu criador, ou mesmo por outros. Bishop (1999) examinou como os pesquisadores podem manipular artigos de periódicos eletrônicos (por exemplo) para criar novos documentos. Ela se refere a isso como desagregação (o desmembramento do artigo) e reagregação (reunir todos os pedaços do artigo, ou parte deles, numa organização diferente). Ademais, alguns documentos da Rede são 'virtuais' — documentos "que carecem de um estado de permanência" (são criados a caminho do usuário) (Watters, 1999). Giordano (2000) salienta que:

> [...] a própria estrutura do documento é problemática porque, num ambiente que se baseia na Rede, um documento que apareça na estação de trabalho de um usuário

> como um único objeto poderá de fato ser uma montagem de documentos vinculados, mas independentes, residentes em bases de dados distribuídas (p. 243).

Essa situação fluida gera confusão em pessoas habituadas ao ambiente bastante sólido e permanente da impressão em papel, porém nem sempre apresentará problemas de indexação e redação de resumos. As mudanças que o autor fizer num texto 'autorizado' exigirão, naturalmente, algumas alterações num resumo ou termos de indexação relativos a esse texto (por exemplo, em portais que apontem para ele). O documento 'virtual' (descrito por Watters) somente estaria qualificado para ser indexado ou resumido se fosse capturado e armazenado numa base de dados como um item novo. Do mesmo modo, o documento reagregado (descrito por Bishop) provavelmente seria um documento informal, que não mereceria os cuidados da indexação e resumo. A impermanência dos documentos eletrônicos tem mais probabilidade de constituir um problema nas intranets de empresas, onde os documentos podem desaparecer por completo, ser radicalmente alterados ou agregados/desagregados sem qualquer controle.

Abordagens profissionais

Duas importantes abordagens que oferecem acesso intelectual aos recursos mais importantes da Rede já estão disponíveis e foram focalizadas no capítulo anterior: a iniciativa CORC (Cooperative Online Resource Catalog) (renomeada 'Connexion' em 2002) e vários portais especializados. Embora a maioria dos portais hajam sido desenvolvidos em áreas 'acadêmicas', a importância deste tipo de atividade para a biblioteca pública foi assim realçada por Holt (1995):

> [...] o pessoal de biblioteca pública pode poupar o tempo de seus clientes ao organizar a massa de informações eletrônicas disponíveis em servidores locais, nacionais e internacionais [...] [e] pode desenvolver guias eletrônicos que ajudem os consulentes em meio aos metadados e megaarquivos, em linha, com que lidarão (p. 555-556).

Ele menciona, especificamente, a importância de proporcionar anotações aos usuários, e encara a biblioteca pública como uma central de informações provida de 'agentes de informação'.

Todas essas atividades dizem respeito à filtragem de recursos da Rede e todas implicam alguma forma de provisão de acesso por assuntos por meio de indexação ou classificação, e talvez alguma forma de resumo. Trippe (2001) ressalta a necessidade de mais classificação dos recursos da Rede, e Elrod (2000) resume um debate em linha sobre a conveniência de as bibliotecas atribuírem números de classificação aos recursos da Rede aos quais elas proporcionam acesso (algumas já o fazem).

Vários autores (ver, por exemplo, MacDougall, 2000, e Studwell, 2000) insistem no uso de vocabulários controlados na indexação de recursos da Rede, porém se mantêm vagos quanto à aplicação ou parecem subestimar grandemente os problemas da aplicação.

Anderson e Perez-Carballo (2001) argumentam que o tremendo aumento na quantidade de texto indexável, especialmente na Rede, torna essencial uma abordagem seletiva da indexação por seres humanos:

> O que não podemos nos permitir é continuar tratando todos os documentos que ingressam em nossos acervos e bases de dados de recuperação da informação como se fossem igualmente importantes e merecedores por igual do nosso trabalho especializado de análise e indexação. Simplesmente, eles não são, e a continuar assim estaremos desperdiçando nossos preciosos recursos (p. 274).

E fazem sugestões sobre como identificar esses itens de escol.

Abordagens alternativas

Drott (2002) propõe uma solução completamente diferente. Ele chamou atenção para o problema da indexação na Rede da seguinte forma:

> Localizar informações sobre temas específicos na Rede é difícil e fica cada vez mais difícil. Os recentes avanços em buscas automáticas na Rede e indexação algorítmica foram grandemente superados pelo enorme crescimento da quantidade de material disponível. As estimativas de cobertura pelos mecanismos de buscas da Rede, feitas por Lawrence e Giles (1999), sugerem a impossibilidade de empregar robôs para indexar toda a Rede, e, evidentemente, quanto maior for o tempo de análise que um robô dedicar à extração de termos de indexação para uma única página, menor será a quantidade do material disponível que poderá ser indexada. Além disso, embora grandes progressos estejam sendo feitos para melhorar a exatidão da indexação automática, ainda é verdade que atribuir termos de indexação a uma base de dados tão diversa quanto a Rede continua sendo um problema para o qual há poucas soluções promissoras (p. 209-210).

E sugere, no entanto, que embora o emprego de indexadores profissionais não seja uma proposta economicamente atraente, os responsáveis pela criação de páginas na Rede deveriam ter condições de eles mesmos fazerem um trabalho aceitável de indexação:

> Seria bom estimular os criadores de sítios na Rede a atribuir seus próprios termos de indexação? O atual modelo de indexação, como o que se encontra nos principais serviços de indexação de periódicos, baseia-se no emprego de indexadores capacitados e que receberam extenso treinamento. Encontra-se, contudo, uma pesquisa encorajadora de Coombs (1998) sobre a indexação de páginas do governo do estado de Washington na Rede. Coombs valeu-se, como indexadores, das pessoas que criaram e trabalharam com os documentos. Os resultados desse estudo mostraram que, quando os indexadores leigos compartilham o mesmo entendimento quanto ao conteúdo e usos de seus documentos, as palavras-chave que produzem são um auxílio razoável na localização por assuntos (p. 218).

E, finalmente:

> Nosso modelo de indexação da Rede bem que poderia tornar-se um desses modelos de 'caos global, ordem local' em que a indexação de campos específicos feita pelo

> autor é adequada dentro de campos limitados, mas ruim para integrar-se em qualquer esquema global de conhecimento. Este conceito sugere um sistema de indexação de duas camadas em que o processamento distribuído de metaetiquetas por um grande número de computadores que rodem programas bastante simples é suportado no nível seguinte por robôs de indexação mais complexa. Esses robôs serão projetados não para extrair de cada página descrições de conteúdo específicas, mas para se concentrar na colocação de grupos de páginas ou sítios inteiros em categorias de assuntos específicos e deixando as informações de conteúdo para os criadores das etiquetas (p. 218).

Outra possibilidade é promover a indexação de recursos da Rede por seus usuários. Besser (1997) analisou a necessidade disso. Embora lidasse especificamente com imagens na Rede, o método é aplicável a quaisquer recursos:

> Se pudermos desenvolver sistemas para terminologia atribuída pelo usuário, os gerentes de acervos poderão apoiar-se nos usuários para que atribuam termos ou palavras-chave a cada imagem. Nesse sistema, quando o usuário encontrasse uma imagem, o sistema lhe perguntaria quais as palavras que teria usado para buscar essa imagem. Essas palavras seriam então inseridas no sistema de recuperação, e usuários subseqüentes que fizessem buscas com essas palavras encontrariam a imagem. À medida que crescer a quantidade de pessoas que usarem esse sistema, também crescerá a quantidade de pontos de acesso para muitas imagens.
>
> É essencial que esses sistemas permitam a realização de buscas em termos atribuídos de forma oficial, tanto independentemente dos termos contribuídos pelos usuários quanto junto com eles. Podemos ter dois tipos de buscas: uma que somente examina termos atribuídos por catalogadores, e a outra que examina tanto os termos atribuídos pelos catalogadores quanto os termos atribuídos pelos usuários. Sistemas desse tipo também poderão servir como auxílio aos catalogadores. Pode-se imaginar um sistema em que, de tempos em tempos, termos contribuídos pelos usuários sejam 'promovidos' à condição de termos oficialmente atribuídos pelo catalogador (e serão então recuperáveis por ambos os métodos).
>
> À medida que sistemas como esse crescem, os usuários futuros poderão querer limitar suas buscas a termos atribuídos por pessoas em quem confiam (talvez porque provenham do mesmo campo ou porque atribuam termos de modo mais confiável). Portanto, provavelmente esses sistemas desenvolverão tanto uma característica pesquisável de 'propriedade' para cada termo atribuído e um 'nível de confiança' que o usuário pode definir e que se aplica a um grupo de proprietários. O projeto de sistemas como este terá também de ser sensível à privacidade de quem contribui com termos. Os usuários que definem níveis de confiança para os atribuidores de termos podem localizar essas pessoas por meio de perfis básicos de sua especialidade e cargo (mas sem identificação), ou podem localizá-los ao encontrar correlações entre outros atribuidores de termos e como o próprio usuário atribui termos a outras imagens [...] (p. 24-25).

A indexação de documentos da Rede feita pelos usuários também foi defendida por Villarroel et al. (2002).

Abordagens automáticas

Encontram-se disponíveis programas que fazem automaticamente a indexação ou resumos de recursos da Rede. Jacsó (2002) avalia alguns programas de sumarização disponíveis no comércio, e Reamy (2002) refere-se a programas de 'autocategorização' (isto é, que colocam automaticamente os recursos em categorias) e prevê importantes avanços nessa área no futuro. A situação do desenvolvimento de métodos automáticos foi examinada no capítulo 15.

Conclusão

Depois de tudo isso, pode-se concluir que as atividades de indexação e resumos vêm aumentando ao invés de diminuir de importância, e que os profissionais dessas áreas podem dar uma contribuição substancial seja no nível de um sítio da Rede ou em níveis mais amplos, como o projeto e implementação de um portal.

Poderão também desempenhar importantes papéis na operação de intranets de empresas. De fato, Reamy (2002), especialista na área de gestão do conhecimento, embora preveja o crescimento da 'autocategorização', oferece enfática defesa da necessidade de profissionais em atividades de acesso intelectual:

> As empresas não querem pagar aos bibliotecários para categorizar seu conteúdo porque acham que sai muito caro. Estão erradas, pelo menos quando se computa o tempo que os funcionários desperdiçam ao tentar em vão encontrar aquele documento de que precisam para responder aquela pergunta do cliente, sem o que o cliente irá embora em busca de um concorrente que, ao contrário, tem a resposta. Apesar disso, muitas empresas ainda não pagarão para que seres humanos categorizem seu conteúdo, e é mais provável que estejam dispostas a pagar entre 250K a 750K por um programa de computador que amiúde executa um trabalho menos eficaz (p. 18).

E acrescenta:

> Em primeiríssimo lugar, a autocategorização não pode substituir por completo um bibliotecário ou arquiteto de informação, embora possa torná-los mais produtivos, poupar seu tempo e produzir um melhor produto final. O próprio programa, sem uma categorização baseada em regras feitas por seres humanos, não pode atualmente chegar a mais de uns 90% de exatidão — o que soa muito bem até se perceber que um de cada dez documentos listados nos resultados de uma busca ou interface de pesquisa estará errado. E, o que é mais importante, estará errado por razões inexplicáveis — razões que levarão os usuários a perder confiança no sistema.
>
> Embora seja muito mais rápida do que um categorizador humano e não exija férias nem plano de saúde, a autocategorização simplesmente ainda não é tão boa quanto um categorizador humano. Não pode compreender as sutilezas de significado nem sumarizar como um ser humano porque não comprende coisas como o significado implícito num documento e porque não leva para a tarefa de categorização os contextos significativos que as pessoas levam. Uma coisa que os trabalhos iniciais da IA nos ensinaram é que embora a velocidade seja importante, a velocidade sozinha não pode compensar a falta de compreensão do significado (p. 21).

E finalmente:

> Ao invés de um risco para os profissionais da informação, a autocategorização pode, de fato, não só aprimorar sua capacidade de solucionar problemas de informação do usuário, mas até elevar seu *status* para algo próximo do nível em que deveria estar. Não apenas os bibliotecários e arquitetos da informação produzirão mais e com mais economia, mas terão *software* caro relativo a essa tarefa e, como todos sabemos, nas empresas de hoje, a menos que haja programas caros envolvidos, ninguém pensará que você é útil.
>
> Bem, está certo, talvez haja um pouco de exagero nisso, mas o programa de autocategorização tem o potencial de realçar o que já devia estar claro — que o profissional da informação está empenhado numa atividade fundamental de infra-estrutura. Os profissionais da informação estão ou deveriam estar envolvidos na criação e manutenção da infra-estrutura intelectual de sua instituição. Embora a tecnologia e as infra-estruturas organizacionais hajam merecido mais atenção e recursos, parte do desequilíbrio poderia ser corrigido com a utilização e integração inteligentes de novos programas, novos métodos de trabalho tanto com os provedores de conteúdo quanto com os consumidores de conteúdo, e novas formas de apresentar a informação.
>
> Portanto, como conclusão, acho ser provável que a autocategorização, em última análise, melhorará tanto o poder quanto o prestígio do profissional da informação (p. 22).

Parece claro que o crescimento continuado dos recursos de informação acessíveis em rede fará com que as atividades de análise temática venham a ter uma importância maior do que jamais tiveram. Além disso, é provável que mais e mais indivíduos estarão envolvidos nessas funções. Com certeza, os métodos para elaboração automática de índices e resumos continuarão a melhorar. No entanto, como Lancaster e Warner (2001) salientam na revisão que escreveram sobre esta área, provavelmente ainda decorrerá muito tempo até que as máquinas sejam suficientemente inteligentes para substituir por completo os seres humanos nessas importantes atividades, se é que de fato um dia o farão.

Parte 2

Prática

Capítulo 18

Exercícios de indexação

Fazendo é que se aprende, seja na indexação e redação de resumos seja em outras atividades. Os dois últimos capítulos deste livro contêm alguns exercícios de indexação e redação de resumos. Evidentemente, os poucos exercícios que podem ser incluídos num livro deste tipo estão longe do que seria suficiente para formar indexadores e resumidores consumados. Apesar disso, apresentamo-los na esperança de que pelo menos proporcionarão alguns exemplos concretos dos principais pontos mencionados nos capítulos precedentes.

Nas poucas páginas a seguir encontram-se vários resumos de relatórios ou artigos de periódicos. Alguns são resumos verdadeiros de publicações existentes. Outros são de artigos 'hipotéticos', embora se baseiem em publicações existentes.

Você deverá indexar cada um desses itens empregando termos do *UNBIS thesaurus* (New York, United Nations, Dag Hammarskjöld Library, 1995).* Se assim lhe aprouver, você poderá primeiro escrever as palavras ou expressões que representem sua análise conceitual de cada item e, em seguida, procurar traduzir cada um desses enunciados para um termo ou termos do tesauro. De qualquer modo, separe seus descritores em descritores *principais* e *secundários*, sendo os primeiros os termos que você considera mais importantes para representar o conteúdo temático.

Depois dos resumos você encontrará nossas sugestões de indexação para cada item, o que lhe permitirá comparar sua indexação com a minha. Lembre-se, contudo, que a indexação é um processo algo subjetivo. Embora acredite na minha indexação, não posso garantir que ela seja 'correta' em sentido absoluto. Incluem-se explicações sobre por que a indexação foi feita de determinada forma.

Os itens 6-13 foram publicados originalmente no número de janeiro de 1977 de *A.I.D. Research and Development Abstracts* e são aqui reproduzidos com permissão do Center for Development Information and Evaluation, United States Agency for International Development.

Itens a serem indexados

1. *O álcool combustível hoje* [Alcohol fuel today] (Baseado em artigo publicado em *Smithsonian*, March 1981, p. 44-53)

---

* Um substituto parcial deste tesauro, em português, é o *Tesauro SPINES* (Brasília: IBICT; Lisboa: INICT, 1988), onde se encontram mais de 70% dos descritores das respostas dadas pelo autor. (N.T.)

Descreve as várias fontes das quais se pode destilar etanol, abrangendo diversos tipos de produtos e resíduos agrícolas, além de resíduos urbanos e lama industrial. Compara os custos de produção do etanol com os da gasolina, e analisa os problemas inerentes à conversão da produção de etanol da fase de usina-piloto à produção comercial em larga escala. Examina as vantagens e desvantagens do gasool, uma mistura de gasolina e álcool combustível, e estuda os problemas que devem ser resolvidos para que os carros a álcool se tornem viáveis.

2. *A erosão e o agricultor.*

Descreve como o vento, a chuva e a neve derretida podem erodir valiosas terras de cultivo, e avalia o volume das perdas agrícolas devidas a essas causas na Europa setentrional. Examina possíveis soluções, a saber, a rotação da cultura de grãos com a de gramíneas protetoras do solo e o emprego de árvores e terraços como quebra-ventos.

3. *A fotografia aérea e o que ela pode fazer* [Aerial photography and what it can do] (Baseado em artigo publicado em *Smithsonian*, March 1984, p. 150-155.)

Faz uma revisão dos vários usos possíveis da fotografia aérea, que abrangem a fotografia por satélite, a vigilância militar, o controle do desarmamento, o estudo de sítios arqueológicos, aplicações em censos (por exemplo, contagem de domicílios), previsão do tempo e inundações, e cartografia (fotogrametria).

4. *O fim do bordo sacarino?* [The end of the sugar maple?] (Baseado em artigos publicados em *Blair & Ketchum's Country Journal*, March 1986, p. 46-49 e *American Forests*, November-December 1987, p. 26-34.)

Uma grande quantidade de árvores de bordo sacarino, no Canadá e norte dos Estados Unidos, ou estão morrendo, ou já morreram, causando uma grave redução na produção de açúcar. Suspeita-se que a principal causa disso seja a chuva ácida que provoca o desfolhamento.

5. *Poderá um avião voar eternamente?* [Can a plane fly forever?] (Baseado em artigo publicado em *Newsweek*, September 28, 1987, p. 42, 47.)

Será testado no Canadá o protótipo de uma aeronave movida a eletricidade, que não precisa de qualquer combustível convencional. A eletricidade é transmitida do solo sob a forma de energia de microondas, sendo reconvertida em eletricidade por meio de 'retenas' instaladas no avião. Teoricamente, o avião poderia permanecer no ar durante meses sem piloto. Entre suas aplicações incluem-se pesquisa científica, vigilância (militar, policial ou civil), previsão do tempo e transporte de passageiros. As microondas também podem acionar espaçonaves. Os possíveis riscos das microondas para a saúde seriam um obstáculo a sua ampla aplicação.

6. *Educação nutricional em programas de alimentação infantil nos países em desenvolvimento.* [Nutrition education in child feeding programs in the developing countries] (Agency for International Development 1974, 44p.)

Este folheto, de texto simples, complementado com desenhos de inúmeros cartazes, destina-se a agentes comunitários e outras pessoas que lidam com a alimentação de crianças nos países em desenvolvimento, ajudando-os a transmitir a mães e filhos ensinamentos sobre os alimentos de que as crianças precisam durante o crescimento e para manter a saúde, e como utilizar alimentos locais na melhoria de sua dieta. Os capítulos abrangem: A dupla finalidade dos programas de alimentação infantil; O que você deve saber sobre os alimentos; Fixando metas que se ajustem a sua comunidade; Algumas regras gerais para o ensino; Trabalhando com mães de pré-escolares; e Ensinando às crianças em programas de alimentação escolar. Percebe-se que a educação nutricional ministrada por pessoas da própria comunidade tem efeito mais duradouro e contribui para a prevenção da desnutrição tanto quanto os alimentos doados, por mais importantes que estes sejam para a saúde das mães e crianças que os recebem.

7. *Melhoramento da qualidade nutritiva e da produtividade da cevada para regiões semi-áridas; relatório anual, 1975/1976.* [Improvement of the nutritive quality and productivity of barley for semi-arid regions; annual report, 1975/1976] (Montana State University, College of Agriculture 1976, 70p.)

Este é o segundo relatório anual de um projeto trienal destinado a melhorar o valor nutritivo de cevada consumida em países menos desenvolvidos, e aumentar a produtividade e diminuir as perdas causadas por doenças da cevada. Durante o primeiro ano de trabalho, foram visitados vários países menos desenvolvidos, a fim de estabelecer contatos e coletar amostras de organismos causadores das principais doenças. O estudo sobre as doenças evoluiu ao ponto de permitir o início de um importante programa de extensão. Quanto ao estudo sobre valor nutritivo, aperfeiçoou-se a técnica de prova microbiológica para determinação de lisina até se obter um instrumento de seleção confiável. Não se verificaram diferenças significativas no valor nutritivo de pares isogênicos Compana, glutinosos e normais, devidas ao tipo de amido ou de composição em aminoácidos da proteína. Os resultados preliminares indicam que os povos que consomem basicamente arroz dariam preferência e provavelmente consumiriam mais uma cevada de endosperma glutinoso do que a cevada de endosperma normal. A variedade High Amylose Glacier apresentou um valor energético levemente menor do que a Glacier normal, mas a primeira contém uma proteína de melhor qualidade, devido a um aumento na proteína de vários aminoácidos. Os dados de desempenho animal (Crescimento, DES e VB) confirmam as análises químicas da composição em proteína e aminoácidos das cevadas Hyproly e Normal Hyproly. Verificou-se que o conteúdo de lisina da proteína era influenciado pelo meio ambiente de modo diferencial, dependendo do gene presente, e se refletia no desempenho animal. O desempenho animal tem alta correlação com o conteúdo de aminoácidos essenciais das cevadas. Comumente a lisina responde por mais de 50% da variação animal em crescimento e PER, e 60% da variação em valor biológico. Identificou-se uma translocação dupla que será

eficaz na transferência do gene Hiproly para uma população, bem como genes resistentes a doenças (cochonilha, ferrugem, nanismo da cevada amarela) no cromossomo 3. Foram desenvolvidas linhagens férteis, cheias e com alto teor de lisina, a partir de híbridos Hiproly, para servirem como matrizes em outros trabalhos de desenvolvimento de variedade com esse gene.

8. *Mulheres africanas no desenvolvimento agrícola, um estudo de caso em Serra Leoa.* [African women in agricultural development, a case study in Sierra Leone] (Spencer, D.S.C., 1976, 41p. Department of Agricultural Economics, Michigan State University.)

Estudo sobre as conseqüências na mão-de-obra, entre famílias de agricultores, de um empréstimo para projeto da A.I.D. destinado ao desenvolvimento de terras alagadas no interior para produção de arroz. O estudo correspondia a uma pequena parte de uma pesquisa nacional sobre problemas de emprego rural em Serra Leoa. Uma aldeia, Benduma, numa das três áreas operacionais do projeto da A.I.D., foi selecionada para o estudo intensivo do trabalho diário realizado por homens, mulheres e crianças, em 23 domicílios selecionados. De maio de 1974 a junho de 1975, foram feitas entrevistas duas vezes por semana em domicílios selecionados, aplicando-se um questionário de insumo–produto para manter registros diários de horas trabalhadas por membro da família e produto não-agrícola, vendas agrícolas e não-agrícolas, empréstimos concedidos e recebidos, e presentes dados e recebidos. A partir desses dados calcularam-se a receita domiciliar por fonte e sua distribuição, utilização da mão-de-obra, rendimentos da mão-de-obra, e perfis sazonais de empresas agrícolas e não-agrícolas. O autor conclui que as mulheres envolvidas no projeto de desenvolvimento trabalharam com um pouco mais de afinco do que as mulheres que dele não participaram, mas que o aumento da carga de trabalho foi muito menor do que o aumento da carga de trabalho de homens adultos e crianças. As mulheres desempenham papel substancial no cultivo de um produto de 'desenvolvimento' (arroz irrigado) que emprega tecnologia aperfeiçoada. Todavia, os resultados do estudo negam a hipótese de que esses projetos de desenvolvimento agrícola impõem uma carga de trabalho desigual para as mulheres em comparação com os homens.

9. *Política de ciência e tecnologia, administração e planejamento da pesquisa na República Árabe do Egito 1976*, 103 p. [Science and technology policy, research management and planning in the Arab Republic of Egypt] (National Academy of Sciences, National Research Council, Washington, D.C. 20418.)

Relatório de simpósio sobre planejamento de política científica e oficina sobre administração e planejamento da pesquisa. A conferência girou em torno de planejamento e política de ciência e tecnologia, e administração da pesquisa. Escolheu-se o formato de 'oficina' como o melhor método para reunir um grupo representativo de cientistas da área das ciências físicas, naturais e sociais, economistas, engenheiros e planejadores de desenvolvimento egípcios e norte-americanos. Constatou-se que, embora o Egito careça de uma política científica

nacional formalmente enunciada, as diversas instituições voltadas para a ciência ali criadas e os recursos financeiros alocados para a pesquisa e a educação em ciência constituem uma importante política nacional implícita. A administração desse grande e complexo conjunto de instituições é uma tarefa formidável, e devem ser encetados todos os esforços para garantir sua eficácia e eficiência. A administração da pesquisa nas universidades é um problema muito diferente da administração de institutos de pesquisa aplicada, e deve ser resolvido tão rapidamente quanto possível. Ainda que o programa de pesquisa aplicada do Egito seja um empreendimento de vulto, sua execução provavelmente esteja a necessitar de uma ampla reestruturação e redirecionamento, para ser totalmente eficaz. A transferência de tecnologia à indústria egípcia por outras nações foi e continuará sendo um elemento fundamental no desenvolvimento industrial do Egito. Visando a assegurar uma transferência eficaz de tecnologia e a reduzir seus custos, deveria haver revisões apropriadas da legislação e das práticas nacionais.

10. *Utilização para o consumo humano de espécies marinhas subutilizadas.* [Utilization of underutilized marine species for human consumption] (Constantinides, S.M.; Figueroa, Jose; Kaplan, Harvey, 1974, 11p. International Center for Marine Resource Development, University of Rhode Island.)

Numa época em que os preços do pescado estão em alta e a desnutrição protéica prevalece em muitos países em desenvolvimento, os pescadores, em todo o mundo, devolvem ao mar, para morrer, milhões de toneladas de peixes ricos em proteína. Esses peixes são devolvidos ao mar porque são considerados 'lixo' ou 'refugo', ou são espécies pouco conhecidas sem qualquer valor econômico. Nos Estados Unidos, os pescadores devolvem até 70% dos peixes apanhados nas redes durante a pesca de outras espécies de valor econômico, como linguados e camarões. O homem não pode mais se permitir ignorar as espécies marinhas ricas em proteína. É preciso criar mercados para as espécies subutilizadas, expandindo-os como soluções alternativas em face da queda no abastecimento de espécies comercialmente consolidadas, ampliando desse modo a indústria, estimulando a preservação dos recursos e revitalizando as espécies conhecidas e há muito exploradas. Pode-se lançar mão de soluções convencionais e não-convencionais, a fim de utilizar essas espécies consideradas refugo. A utilização delas pode ser desenvolvida conforme as seguintes atividades principais: carne moída (espécies mistas ou uma única), pastas e produtos secos. A produção de carne moída a partir de inúmeros peixes de tamanho pequeno e médio torna-se viável com o emprego de separadores que produzem carne sem espinhas. A carne é lavada e depois congelada em blocos. Uma combinação de peixes gordurosos e não-gordurosos resulta num produto final apetecível ao consumidor. Pastas de peixe, de camarão e de caranguejo são preparadas segundo vários métodos. Acrescentam-se à carne moída lavada sal, amido e polifosfatos, a fim de produzir uma pasta com a qual podem ser preparados embutidos e outros produtos. Podem também ser elaborados outros produtos como peixe misturado

com batata, pasta para espalhar sobre pão, molhos para salgadinhos, e sopas ou vários tipos de carne moída misturados entre si ou com outros ingredientes para obter novos sabores apetitosos. Espécies que até hoje não foram utilizadas pelo homem serão utilizadas no futuro, e as chamadas espécies sem valor serão aceitas como espécies comestíveis apropriadas ao consumo humano direto.

11. *A utilização de alunos monitores e instrução programada pelo rádio: alternativas viáveis na educação.* [The use of peer tutoring and programmed radio instruction: viable alternatives in education] (Hannum, W.H.; Morgan, R.M. 1974, 38p. Florida State University, College of Education.)

Os educadores de países em desenvolvimento provavelmente obterão melhores resultados ao aplicar os princípios e não os equipamentos da tecnologia educacional. Já foi demonstrado que os princípios do ensino programado são eficazes na promoção da aprendizagem em circunstâncias muito variadas. Os materiais instrucionais mais eficazes podem ser desenvolvidos com a aplicação dos princípios da instrução programada e do aprendizado com proficiência. O rádio, quando combinado com alunos monitores, pode ser um instrumento educacional eficaz em países em desenvolvimento. Os conceitos de ensino programado e aprendizado com proficiência podem ser incorporados ao projeto de programas educacionais pelo rádio. Estes, acompanhados por alunos monitores, aperfeiçoam o esforço educacional global de modo compatível com os recursos de muitos países em desenvolvimento. Este tipo de sistema educacional é uma alternativa viável à educação formal tradicional. Deveria ser testado em vários países em desenvolvimento visando à exploração de todo seu potencial.

12. *Fatores culturais e sociais que influem na participação de pequenos agricultores em programas formais de crédito* [Cultural and social factors affecting small farmer participation in formal credit programs] (Gillette, Cynthia; Uphoff, Norman 1973, 40p. Rural Development Committee, Center for International Studies, Cornell University.)

Este trabalho contém três pressupostos básicos que, com uma exceção, constituem seu tema principal. A exceção é a questão da 'racionalidade econômica', conhecida de todos os que se interessam pelo desenvolvimento do Terceiro Mundo, mas que é vista como justificativa de uma breve análise na introdução. A parte II trata do contexto cultural de pequenos agricultores como tomadores de empréstimo, isto é, diversos fatores que influem sobre a demanda de crédito. Em seguida, a parte III trata do contexto cultural dos programas creditícios como emprestadores, isto é, fatores que condicionam o fornecimento de crédito disponível em termos funcionais aos pequenos agricultores. A parte IV mostra várias implicações das partes II e III: o que acontece quando esses dois sistemas culturais interagem e quais os prováveis pontos de dificuldade. A parte V conclui fazendo uma comparação das diferenças gerais entre fontes de crédito formais e informais.

13. *Desenvolvimento de coberturas de casas de baixo custo a partir de materiais locais em nações em desenvolvimento; relatório anual, 1974/1975* [Development of low-cost roofing from indigenous materials in developing nations; annual report, 1974/1975] (Monsanto Research Corporation, Dayton, Ohio, 1975, 335 p.)

Este relatório examina a segunda fase (maio de 1974 a setembro de 1975) de uma pesquisa de três fases, com três anos e meio de duração, visando à obtenção de melhores coberturas de casas para países em desenvolvimento, mediante a combinação de fibras e enchimentos locais com aglutinantes de baixo custo. A meta final do programa é tornar disponível em pelo menos três países, cada um deles na América Latina, Ásia e África, um sistema de cobertura de casas que seja econômica e tecnicamente aceitável e que dependa menos de divisas estrangeiras do que as alternativas ora existentes. O objetivo do programa será demonstrado, em cada um dos países participantes, com a construção de pelo menos quatro protótipos de coberturas e a transferência da tecnologia necessária a instituições qualificadas. Os países colaboradores atuais são Jamaica, Filipinas e Gana. A prioridade do projeto durante a fase III consistiu no desenvolvimento de materiais de cobertura e estabelecimento do mecanismo de transferência de tecnologia. Os objetivos predominantes do desenvolvimento de materiais incluíam o estabelecimento de um conjunto generalizado de critérios para coberturas; definição dos componentes do material composto; determinação dos conjuntos mais promissores de materiais, processos e produtos; e análises de custos e viabilidade dos sistemas propostos. Foram definidos quatro sistemas propostos de material composto para coberturas que empregam de 70 a 100% de materiais locais. Excepcional como enchimento é o bagaço, que é o resíduo da cana-de-açúcar. Os principais aglutinantes propostos incluem borracha natural, resinas fenólicas e termoplásticas comerciais. A cura acelerada e ao ar livre demonstra a viabilidade dos sistemas propostos. Os objetivos dos aspectos relativos à transferência de tecnologia incluíram a definição de instituições colaboradoras potenciais e pessoas físicas na Jamaica, Filipinas e Gana; a formação de comissões de trabalho, assessoras e técnicas, em cada um desses países que participariam do programa de desenvolvimento de coberturas de casas; e a identificação de instituições qualificadas interessadas na futura produção comercial dessas coberturas. Essas instituições, comissões e grupos de trabalho foram definidos nos três países e estão funcionando, em vários graus, com a Jamaica à frente. Foram identificadas indústrias do setor privado que poderão tornar-se futuros fabricantes de coberturas em cada um dos três países. Durante a fase III, de outubro de 1975 a dezembro de 1976, o programa será concluído com a otimização de materiais, projeto, fabricação, testes e avaliação dos protótipos de coberturas; e fabricação em campo, instalação e avaliação das coberturas em escala integral.

Indexação e explicações do autor
(Os descritores principais são identificados com um asterisco *)

1. *O álcool combustível hoje*
   Alcohol fuels* [álcoois combustíveis]
   Gasohol* [gasool]
   Production costs [custos de produção]
   Gasoline [gasolina]
   Crops [produtos agrícolas]
   Agricultural wastes [resíduos agrícolas]
   Refuse derived fuels [combustíveis derivados do lixo]
   Domestic wastes [resíduos domésticos]
   Industrial wastes [resíduos industriais]
   Pilot projects [projetos-piloto]
   Waste utilization [utilização de resíduos]

Não se encontra no UNBIS o termo *ethanol* [etanol] do qual se faz remissiva para *alcohol fuels* [álcoois combustíveis], que parece ser o termo mais pertinente para este item. Se o termo *ethanol* existisse no tesauro, ele seria usado, e não *alcohol fuels*, apesar do título, pois o resumo indica que o artigo trata exclusivamente de etanol. Não confie em demasia nos títulos; eles às vezes são enganosos.

O resumo sugere que o artigo dá bastante atenção ao gasool, e por isso este termo é também empregado na indexação seletiva. O tesauro não contempla a possibilidade de se expressar a idéia de 'carros a álcool'. No entanto, isso se acha implícito com muita nitidez em *gasohol* [gasool], de modo que o emprego do termo *automobiles* [automóveis], embora não seja errado, parece desnecessário. Se utilizássemos o termo *motor fuels* [combustíveis para motores] estaríamos cometendo um sério engano, porque o artigo trata exclusivamente de gasool, que é um tipo de combustível para motores, e *motor fuels*, no UNBIS, é um termo genérico (TG) de uma ordem superior a *gasohol*.

Na indexação mais exaustiva seria preciso abarcar as outras idéias condensadas no resumo. As fontes do etanol podem ser satisfatoriamente abrangidas por intermédio do termo *crops* [produtos agrícolas] junto com diversos termos específicos de '*waste*' [resíduos]. Uma vez que se mencionam tipos específicos de resíduos, é melhor empregar os termos específicos e não o mais genérico *wastes*. Para exemplificar, suponhamos que alguém estivesse à procura de informações sobre possíveis aplicações de resíduos agrícolas. Este parece ser um item bastante relevante, porém não seria encontrado se estivesse indexado sob o termo mais genérico.

O termo *municipal wastes* [resíduos urbanos] não existe no UNBIS, mas como resíduos urbanos são, em geral, resíduos domésticos (ver nota explicativa em *domestic wastes* no UNBIS), deve-se, por isso, empregar *resíduos domésticos*. Se o artigo der muita atenção ao aspecto 'resíduos', um termo adequado parece ser *waste utilization* [utilização de resíduos]. *Refuse derived fuels* [combustíveis derivados do lixo] é, com certeza, um termo apropriado.

Considerando que se comparam os custos do etanol e da gasolina, o termo *gasoline* deveria ser incluído na indexação exaustiva. *Production costs* [custos de produção] certamente sim.

Os termos do UNBIS não permitem que se expresse com precisão a idéia de 'ampliar' a produção da escala de usina-piloto para a escala comercial. O termo mais pertinente parece ser *pilot projects* [projetos-piloto].

Também é impossível exprimir a idéia de 'vantagens/desvantagens' ou 'problemas' (relativos a carros movidos a gasool ou álcool). A maioria dos vocabulários controlados não chega a contemplar esse tipo de idéias mais nebulosas.

2. *A erosão e o agricultor*
   Soil erosion* [erosão do solo]
   Rain [chuva]
   Soil conservation* [conservação do solo]
   Snow [neve]
   Crop rotation [rotação de culturas]
   Windbreaks [quebra-ventos]
   Crop yields [produção agrícola]
   Europe [Europa]

Aqui, o termo fundamental é *soil erosion* [erosão do solo]. *Soil conservation* [conservação do solo] é o termo que, isoladamente, melhor abrange 'possíveis soluções'. Deficiências do tesauro UNBIS dificultam a indexação exaustiva. *Rain* [chuva], *snow* [neve] e *wind* [vento] são termos apropriados e necessários, caso alguém precise fazer uma busca de artigos especificamente sobre erosão do solo provocada por chuva, neve ou vento. Quanto às soluções específicas analisadas, *crop rotation* [rotação de culturas] e *windbreaks* [quebra-ventos] são apropriados

No UNBIS não se pode expressar a idéia de 'perdas agrícolas', porém *crop yields* [produção agrícola] é suficientemente aproximado para merecer ser atribuído (isto é, o efeito da erosão sobre a produção). O termo *Northern Europe* [Europa setentrional] não existe no UNBIS (embora exista *Southern Europe*! [Europa meridional]), por isso o termo *Europe* deve ser atribuído. Isso exemplifica um aspecto importante: se o termo exato de que se necessita não existe no tesauro, utiliza-se o termo mais específico que o tesauro permite.

3. *A fotografia aérea e o que ela pode fazer*
   Aerial photography* [fotografia aérea]
   Aerial photogrammetry [aerofotogrametria]
   Image analysis [análise de imagens]
   Aerial surveys* [levantamentos aéreos]
   Hydrographic surveys [levantamentos hidrográficos]

Flood control [controle de inundações]
Military reconnaissance [reconhecimento militar]
Satellite monitoring [monitoramento por satélite]
Geodetic satellites* [satélites geodésicos]
Archaeology [arqueologia]
Censuses [censos]
Weather prediction [previsão do tempo]
Weather maps [cartas meteorológicas]

Este artigo parece tratar do emprego de aeronaves e satélites na realização de diversos tipos de levantamentos fotográficos. *Aerial photography* [fotografia aérea] e *aerial surveys* [levantamentos aéreos] são termos importantes. O termo *satellite photography* [fotografia por satélite] não existe no UNBIS. A idéia poderia ser expressa, contudo, combinando-se *aerial photography* com um termo de 'satélite'. O termo mais apropriado parece ser *geodetic satellites* [satélites geodésicos], especialmente porque o UNBIS liga (por meio de TR) o termo *aerial photogrammetry* [aerofotogrametria] com *geodetic satellites*.

Quanto às aplicações, o UNBIS abrange satisfatoriamente umas, e outras não tão satisfatoriamente. *Verification* [verificação] é um termo do tesauro aparentemente apropriado para este artigo até se descobrir que *satellite monitoring* [monitoramento por satélite] é um termo mais específico do que *verification*. Emprega-se *satellite monitoring* porque o tipo de verificação analisado no documento (verificação de desarmamento) só pode ser realizado por meio de fotografias tiradas por satélite. Lembre-se: empregue sempre o termo *mais específico* existente no tesauro, ainda que um outro termo possa 'soar' mais apropriado. Isso exemplifica outro aspecto importante: o 'contexto' de um termo num tesauro revela o significado desse termo, mesmo que não seja acompanhado de uma nota explicativa. O contexto de *satellite monitoring* no UNBIS deixa claro que o objetivo é o uso de satélites na verificação, e não o monitoramento de satélites.*

O estudo de sítios arqueológicos provavelmente fica mais bem abrangido por *archaeology* [arqueologia] do que por *archaeological excavations* [escavações arqueológicas]. Como 'contagem de domicílios' é usado simplesmente como exemplo de uma aplicação em censos, o termo genérico *censuses* [censos] é mais seguro do que *housing censuses* [censos domiciliares]. Além disso, o último termo é um tanto ambíguo, pois pode referir-se aos ocupantes de prédios e não ao número de residências.

No UNBIS, o prognóstico sobre o tempo se traduz como *weather prediction* [previsão do tempo]. Como isso implica a elaboração de cartas meteorológicas, este termo também seria aplicado, ainda que seja um tanto periférico. Não há como abranger a previsão de inundações como tal. O objetivo é a prevenção

* Esta advertência do autor justifica-se por causa da ambigüidade da expressão em inglês, que tanto pode significar 'monitoramento *de* satélite' quanto 'monitoramento *por* satélite'. (N.T.)

contra inundações, de modo que se deve usar *flood control* [controle de inundações]. Como o movimento de água ou gelo se acha implícito, *hydrographic surveys* [levantamentos hidrográficos] também seria um bom termo.

A cartografia está bem abrangida por *aerial photogrammetry* [aerofotogrametria]. Finalmente, como todas as diversas aplicações envolvem extensamente a interpretação de fotografias, *image analysis* [análise de imagens] parece ser inteiramente adequado.

4. *O fim do bordo sacarino?*
Sugar growing* [culturas açucareiras]
Sugar industry [indústria açucareira]
Trees* [árvores]
Defoliation [desfolhamento]
Acid rain* [chuva ácida]
Canada [Canadá]
United States [Estados Unidos]
Plant diseases [doenças das plantas]

O tesauro UNBIS só reconhece como plantas produtoras de açúcar a cana-de-açúcar e a beterraba sacarina, por isso, é preciso empregar aqui *sugar crops* [culturas açucareiras]. Como nesse tesauro há poucos termos para tipos específicos de árvores, é preciso empregar o termo genérico *trees* [árvores]. É provável que a poluição seja a causa do desfolhamento, mas é desnecessário usar *air pollution* [poluição atmosférica] porque *acid rain* [chuva ácida] é mais exato.

5. *Poderá um avião voar eternamente?*
Aircraft* [aeronave]
Electric vehicles* [veículos elétricos]
Microwaves* [microondas]
Scientific research [pesquisa científica]
Prototypes [protótipos]
Spacecraft [espaçonave]
Radiation sickness [doença provocada por radiação]
Military surveillance [vigilância militar]
Canada [Canadá]

A idéia de uma aeronave movida a eletricidade, que utilize microondas, acha-se bem abrangida pelos três termos com asterisco. O artigo concede mais atenção às possíveis aplicações científicas e militares, pelo que se fez um esforço para englobar esses aspectos. Lamentavelmente, a idéia da vigilância em geral está ausente do UNBIS, mas existe *military surveillance* [vigilância militar]. As outras possíveis aplicações mencionadas no artigo, como, por exemplo, a previsão do tempo, o são de modo tão superficial que parecem não merecer sua inclusão na indexação. Como o risco para a saúde mencionado é a radiação de microondas, o termo *radiation sickness* [doença provocada por radiação] parece apropriado.

6. *Educação nutricional em programas de alimentação infantil*
   Child feeding* [alimentação infantil]
   Nutrition education* [educação nutricional]
   Child nutrition* [nutrição infantil]
   Developing countries [países em desenvolvimento]
   Infant nutrition [nutrição do lactente]
   School meals [merendas escolares]
   O assunto desse relatório está perfeitamente abarcado pelos termos existentes no tesauro.

7. *Melhoramento da qualidade nutritiva e da produtividade da cevada*
   Barley* [cevada]
   Arid zones* [zonas áridas]
   Nutrition* [nutrição]
   Crop yields [produção agrícola]
   Developing countries [países em desenvolvimento]
   Plant breeding [melhoramento genético de plantas]
   Plant genetics [genética vegetal]
   Plant diseases [doenças das plantas]
   Plant protection [proteção das plantas]
   Proteins [proteínas]
   Arid zones [zonas áridas] é, no UNBIS, o mais próximo que se pode chegar de 'regiões semi-áridas'.

8. *Mulheres africanas no desenvolvimento agrícola*
   Rice [arroz]
   Sierra Leone [Serra Leoa]
   Women in agriculture* [mulheres na agricultura]
   Women workers* [mulheres trabalhadoras]
   Women in development [mulheres no desenvolvimento]
   Women's rights [direitos das mulheres]
   Hours of work* [horas de trabalho]
   Working time arrangement [organização do tempo de trabalho]
   Labour productivity [produtividade do trabalho]
   Division of labour [divisão do trabalho]
   Não se deixe enganar pelo título. Este documento é sobre mulheres em Serra Leoa, não sobre mulheres africanas em geral. O artigo estuda principalmente as condições de emprego das mulheres, não a cultura do arroz. Embora *rice* [arroz] seja um termo pertinente, os mais importantes são *women workers* [mulheres trabalhadoras] e *hours of work* [horas de trabalho]. *Arroz* não é um termo principal, pois quem estiver à procura de itens sobre a cultura do arroz poderá não se interessar por esse tipo de estudo social. O termo *division of labour* [divisão do trabalho] provavelmente é pertinente, uma vez que o documento analisa a relação homens/mulheres no trabalho, no entanto, a nota explicativa no tesauro traz uma indicação muito inadequada sobre como e quando usar este termo.

9. *Política de ciência e tecnologia*
   Egypt* [Egito]
   Science and technology policy* [política de ciência e tecnologia]
   Science and technology planning* [planejamento de ciência e tecnologia
   Research and development* [pesquisa e desenvolvimento]
   Technology transfer [transferência de tecnologia]
   Scientific research [pesquisa científica]
   Public administration [administração pública]
   Management [administração]
   Science and technology financing [financiamento de ciência e tecnologia]
   São necessários vários termos para abranger esse relatório de modo adequado. Note-se que *research and development* [pesquisa e desenvolvimento] e *management* [administração] são ambos necessários para refletir a idéia de 'administração da pesquisa'. *Egypt* [Egito] é considerado um termo principal porque todo o relatório trata da situação egípcia. O que é muito diferente do artigo sobre 'mulheres africanas', no qual a localização (Serra Leoa) é quase acidental para a finalidade do estudo.

10. *Utilização para o consumo humano de espécies marinhas subutilizadas*
    Food consumption* [consumo de alimentos]
    Fish* [peixes]
    Fish processing [processamento de peixes]
    Fishery products [produtos da pesca]
    Fishery conservation* [preservação da pesca]
    Este é um exemplo de um artigo que não pode ser indexado adequadamente porque o tesauro não expressa a idéia de 'espécies de peixes subaproveitadas'. Os termos aqui empregados não oferecem uma imagem satisfatória daquilo de que trata o item, mas são os melhores existentes.

11. *A utilização de alunos monitores e instrução programada pelo rádio*
    Educational radio* [ensino pelo rádio]
    Programmed instruction* [instrução programada]
    Developing countries [países em desenvolvimento]
    Nonformal education [educação não-formal]
    Teaching personnel [pessoal de ensino]
    Mais uma vez um item que não foi abrangido satisfatoriamente porque o tesauro carece de termos que expressem a idéia de 'alunos monitores'ou mesmo de 'monitoria'.

12. *Fatores culturais e sociais que influem sobre a participação de pequenos agricultures em programas formais de crédito*

Credit policy* [política de crédito]
Farmers* [agricultores]
Small farms* [pequenas propriedades agrícolas]
Developing countries [países em desenvolvimento]
Agricultural credit* [crédito agrícola]
Cultural values [valores culturais]
Social values [valores sociais]

Este é um excelente exemplo de um relatório relativamente longo que é satisfatoriamente abrangido com um pequeno número de termos. Para exprimir a idéia de 'pequenos agricultores' é preciso usar tanto *farmers* [agricultores] quanto *small farms* [pequenas propriedades agrícolas]. Atribui-se *developing countries* [países em desenvolvimento] porque é óbvio que este é o contexto no qual se analisa o crédito agrícola.

13. Desenvolvimento de coberturas de casas de baixo custo

Roofs* [coberturas de casas]
Traditional technology [tecnologia tradicional]
Bagasse [bagaço de cana]
Fibres [fibras]
Building materials* [materiais de construção]
Technology transfer [transferência de tecnologia]
Rubber [borracha]
Plastic products [produtos plásticos]
Jamaica
Ghana [Gana]
Philippines [Filipinas]
Developing countries [países em desenvolvimento]

Esta indexação não é totalmente satisfatória porque o tesauro não nos permite expressar '*indigenous materials*' ['materiais locais']. No entanto, podem-se considerar os materiais locais como relacionados de perto com a tecnologia local, de modo que o termo *traditional technology* [tecnologia tradicional] se justifica, ainda que não seja exatamente o ideal.

Capítulo 19

Exercícios de redação de resumos

Parte 1

Para fazer este exercício é preciso primeiro reunir os artigos de periódicos que são mencionados na lista abaixo. A maioria deles é facilmente encontrada em bibliotecas dos Estados Unidos. Para cada artigo prepare um resumo ou resumos (ver nota adiante) e compare o que você escreveu com os resumos que sugeri e com meus comentários. De que modo esses resumos diferem dos seus? Quais os melhores? Por quê?

Artigos a serem resumidos:

1. Can a plane fly forever? (*Newsweek*, September 28, 1987, p. 42, 47).
2. Pluto: limits on its atmosphere, ice on its moon (*Science News*, September 26, 1987, p. 207).
3. Plastic shocks and visible sparks (*Science News*, September 5, 1987, p. 152).
4. Moscow's chemical candor (*Newsweek*, October 19, 1987, p. 56).
5. Stereotypes: the Arab's image (*World Press Review*, June 1986, p. 39).
6. Ads require sensitivity to Arab culture, religion (*Marketing News*, April 25, 1986, p. 3).
7. France, racism and the Left (*The Nation*, September 28, 1985, p. 279-281).
8. Compassion for animals (*National Forum*, Winter 1986, p. 2-3).

Nota: Para o item 1, redija resumos indicativos. Para os itens 2, 5 e 7, redija resumos informativos. Para os de número 3 e 4, redija resumos indicativos e informativos. Para os itens 6 e 8, faça da forma que lhe parecer mais adequada.

Resumos deste autor

1. **Can a plane fly forever?** [Poderá um avião voar eternamente?] (*Newsweek*, September 28, 1987, p. 42, 47).

Resumo (indicativo)

Será testado no Canadá o protótipo de uma aeronave movida a eletricidade, que não requer combustível convencional. A eletricidade é transmitida do solo sob a forma de energia de microondas e reconvertida em eletricidade por 'retenas' no avião. Teoricamente, o avião pode permanecer no ar durante meses sem piloto. Suas aplicações incluem pesquisa científica, vigilância (militar, policial ou civil), previsão do tempo e transporte de passageiros. As microondas podem

também acionar espaçonaves. Possíveis riscos das microondas para a saúde podem impedir aplicação mais generalizada.

Comentário

A clareza tem precedência sobre a brevidade. A expressão 'que não requer combustível convencional' é necessária para esclarecer que o aparelho é *inteiramente* movido a eletricidade. O resumo não deve extrapolar o que o artigo afirma. Assim, 'será testado' está bem, mesmo que o resumidor saiba que os testes já foram realizados. Procure evitar o emprego de palavras irrelevantes. Por exemplo, 'As microondas podem também acionar espaçonaves' é mais conciso do que 'As microondas podem também ser aplicadas com a finalidade de acionar espaçonaves', sem que com isso se perca em clareza. Como não se apresentam resultados concretos, seria difícil escrever um resumo verdadeiramente informativo desse item.

2. **Pluto: limits on its atmosphere, ice on its moon** [Plutão: limites de sua atmosfera, gelo em sua lua] (*Science News*, September 26, 1987, p. 207).

Resumo (informativo)

Cálculos recentes indicam que o diâmetro de Plutão talvez não supere 2 290 km, com sua lua, Caronte, cujo diâmetro não deve ser superior a 1 284 km. O espectro infravermelho de Plutão parece ser radicalmente diferente do de Caronte. Plutão possui uma superfície rica em metano, mas Caronte, com relativamente pouco metano, parece ter uma predominância de gelo de água. A refletividade de Caronte corresponde somente à metade da de Plutão, sugerindo que Plutão possui uma temperatura superficial mais baixa: talvez 50 kelvin em Plutão e 58 em Caronte. A pressão de vapor em Plutão pode ser de apenas 3,5 microbars comparada com 59 em Caronte. Parece que Plutão possui calotas polares não-estáticas de gelo de metano cuja cobertura do planeta varia com o tempo.

Comentário

Este é um resumo realmente informativo, que procura condensar todos os principais dados descritos no artigo. Procure evitar redundância. Por exemplo, é exato, mas não necessário, dizer 'Medidas do espectro infravermelho sugerem que o espectro infravermelho de Plutão parece ser radicalmente diferente do de Caronte' porque a referência a 'espectro infravermelho' por si mesma indica que foram tomadas medidas do espectro infravermelho.

3. **Plastic shocks and visible sparks** [Choques plásticos e faíscas visíveis] (*Science News*, September 5, 1987, vol. 132, no. 10, p. 152).

Resumo (indicativo)

Descreve as condições sob as quais a eletricidade estática pode causar incêndios ou explosões, ao se manusear pós ou líquidos, e menciona dois instrumentos

desenvolvidos recentemente que podem ser empregados para monitorar operações de manuseio de materiais.

Resumo (informativo)

Ao encher ou esvaziar recipientes, a eletricidade estática pode produzir faíscas que causam incêndios ou explosões. Vasilhames plásticos que contenham líquidos inflamáveis podem receber uma carga vinda de um saco plástico ou do bolso de um casaco que esteja perto, produzindo uma faísca quando o líquido é despejado. As cargas ocorrem também quando do transporte de pós químicos, quando tambores de metal revestidos de plástico são enchidos com líquidos condutores ou recebem trapos embebidos com solventes condutores, ou quando revestimentos semicondutores que tenham solventes como base são aplicados à superfície de uma película não-condutora. O próprio corpo humano pode produzir faíscas que causam ignição de vapores inflamáveis. Novos instrumentos permitem o monitoramento das operações de encher e esvaziar vasilhames com pós ou líquidos. Empregando intensificação eletrônica de imagens ou a medição da polaridade de carga e sua magnitude, registram o faiscamento e identificam as condições mais prováveis de causar a ignição. Os líquidos mais perigosos possuem baixa condutividade, têm carga negativa, são altamente inflamáveis e se evaporam facilmente formando uma mistura de vapor–ar que sustenta a ignição.

Comentário

Eis um bom exemplo da diferença entre resumo indicativo e informativo. O primeiro simplesmente menciona de que trata o artigo, enquanto o segundo procura ser uma síntese verdadeira — quais os tipos de operações, qual o tipo de risco, qual o tipo de instrumento e assim por diante. Muitas vezes consegue-se ser conciso, sem sacrificar a clareza, ao se omitir artigos ou conjunções. Por exemplo, "Ao encher ou esvaziar recipientes..." é mais conciso e tão claro quanto 'Ao encher ou esvaziar os recipientes...'.

4. **Moscow's chemical candor** [A sinceridade química de Moscou] (*Newsweek*, October 19, 1987, p. 56).

Resumo (informativo)

A União Soviética admite abertamente a estocagem de armas químicas, mas afirma que não as produz mais. Foi dada permissão a observadores ocidentais para visitar a base de Shikhani, antes secreta, mas especialistas ocidentais acham que as armas expostas são antigas — os soviéticos teriam armas mais modernas que não admitem possuir. Os EUA afirmam que interromperam a produção de armas químicas em 1969, mas os serviços secretos ocidentais acreditam que os soviéticos ainda as produzem, tendo armazenadas até 300 000 toneladas. Os EUA forneceram minucioso relatório sobre as dimensões e a localização dos estoques norte-americanos, mas os soviéticos se recusam a retribuir isso enquanto não for

assinado um tratado. A proposta norte-americana de eliminação das armas químicas não foi aceita pelos soviéticos em 1984, mas agora afirmam desejar um tratado e a verificação no local. Os soviéticos afirmam que a decisão norte-americana de produzir armas 'binárias' obstruirá a assinatura de um tratado, mas os EUA acham que essa nova geração de armas na realidade forçará os soviéticos a negociar.

Resumo (indicativo)

Descreve medidas adotadas recentemente pela União Soviética para apoiar um tratado de proscrição do emprego de armas químicas. Menciona a nova geração de armas 'binárias' atualmente produzidas pelos EUA e o possível efeito disso na assinatura de um tratado.

Comentário

Mais um bom exemplo da diferença entre resumo indicativo e resumo informativo. O primeiro procura resumir a essência do artigo enquanto o segundo simplesmente indica de que ele trata.

5. **Stereotypes: the Arabs' image** [Estereótipos: a imagem dos árabes] (*World Press Review*, June 1986, p. 39).

Resumo (informativo)

A mídia norte-americana, principalmente a televisão, promove uma imagem negativa dos árabes e dos países árabes. A hostilidade aos árabes, exacerbada pelo conflito árabe-israelense e a crise do petróleo da década de 1970, estende-se a mais de um milhão de árabes que vivem nos Estados Unidos. Os interesses da verdade, da paz e da fraternidade exigem que sejam adotadas medidas para mudar essa imagem.

Comentário

O resumidor deve decidir sobre o que é e o que não é importante. A essência desse breve artigo parece bem abrangida por essas três frases. É dispensável resumir os detalhes sobre os estereótipos, que ocupam cerca de metade do artigo. A inclusão dos nomes de instituições mencionadas no artigo tornaria o resumo muito minucioso.

6. **Ads require sensitivity to Arab culture, religion** [A publicidade exige sensibilidade à cultura e religião árabes] (*Marketing News*, April 25, 1986, p. 3).

Resumo

Devido à queda dos preços do petróleo, é preciso que uma publicidade eficaz estimule os países árabes a consumir. Os publicitários devem compreender os costumes religiosos, sociais e culturais que presidem a vida árabe. Apresentam-se alguns exemplos de coisas a serem evitadas.

Comentário

Apesar de muito sucinto, este é menos um resumo indicativo do que uma tentativa de resumir o que o autor diz, em vez de descrever aquilo de que trata o artigo. Somente a última frase é realmente indicativa. Isso mostra como os resumos podem ser redigidos de modo a combinar elementos informativos e indicativos.

7. **France: racism and the Left** [França: o racismo e a esquerda] (*The Nation*, September 28, 1985, p. 279-281).

Resumo (informativo)

O partido ultradireitista, Frente Nacional, promove ativamente o ódio racial na França, principalmente contra os norte-africanos, mas os comunistas e socialistas pouco têm feito para lutar contra o preconceito racial. As campanhas contra o racismo são organizadas por grupos não-oficiais, principalmente de jovens.

Comentário

Como no exemplo anterior, este resumo é mais informativo do que indicativo. Uma comparação dos resumos 5–7 com os resumos 1–4 mostrará que é mais difícil redigir resumos verdadeiramente informativos em ciências sociais do que nas ciências exatas. Os artigos em ciências sociais tendem a ser mais abstratos e conter menos dados concretos.

8. **Compassion for animals** [Compaixão pelos animais] (*National Forum*, Winter 1986, p. 2-3).

Resumo

O estreito vínculo entre homens e animais, que costumava existir em épocas passadas, foi corroído pelo desenvolvimento urbano e a industrialização, provocando descaso pela vida animal em muitas partes. No entanto, um forte vínculo homem–animal é fundamental para a saúde do indivíduo, da comunidade e da sociedade. Sugere maneiras pelas quais a sociedade poderia melhorar sua sensibilidade e compaixão pelos animais.

Comentário

Mais uma vez, parece bastante apropriado um resumo combinado indicativo/informativo. As primeiras duas frases, ao tentar condensar a mensagem dos autores, são realmente informativas, enquanto a última frase é evidentemente indicativa. O resumo ficaria totalmente informativo se fossem resumidos todos os métodos destinados a despertar compaixão, mencionados na página 3 do artigo, mas eles são tão variados que seria preciso um resumo bastante extenso, o que aparentemente não se justifica em face da brevidade do próprio artigo.

PARTE 2

Reproduzem-se a seguir oito resumos publicados em *Irricab* (abril de 1980, volume 5, número 2), uma publicação de resumos no campo da irrigação editada pelo International Irrigation Information Center [Centro Internacional de Informação sobre Irrigação]. Você encontra algo de errado nesses resumos? Como melhorá-los? Veja, após cada um deles, os comentários deste autor.

Resumos

[Os resumos são aqui reproduzidos com a gentil permissão do International Irrigation Information Center, Bet Dagan, Israel, e Pergamon Press Inc. A seleção destes resumos nessa fonte foi determinada apenas por razões de conveniência e não implica de forma alguma que os resumos de *Irrricab* sejam de qualidade inferior. Com efeito, em geral, são muito bons, sendo difícil encontrar algum que necessite de grandes melhorias.]

1. Anon. (Clarification of highly turbid waters by means of acoustic filters) (Rus) [Clarificação de águas excessivamente barrentas mediante filtros acústicos] *Gidrotekh Melior*, 1977, (9): 98-99

   Descreve-se sucintamente o desenvolvimento de um método de clarificação da água com filtros acústicos. Estudaram-se as características hidráulicas de vários crivos com e sem vibração, e se determinou o coeficiente de resistência de vários crivos. Propõe-se o método para clarificação da água sem o emprego de reagentes químicos.

2. Vaneyan, S.S.; Makoveev, V.P. (Volzhanka side roll sprinkler for irrigation of vegetable crops) (Rus) [Aspersor Volzhanka de rotação lateral para irrigação de culturas de hortaliças] *Gidrotekh Melior*, Mar 1979, (3): 67-68, 1 photo, 2 tab. (All-Union Research Institute for Vegetable Growing, USSR)

   Relatam-se experiências com a irrigação de culturas de várias hortaliças empregando o aspersor Volzhanka. O artigo contém uma equação para calcular a duração da irrigação e o número de unidades de aspersores necessários para irrigar determinada área. Apresenta dados sobre danos causados aos plantios pelas rodas dos aspersores.

3. Rhoades, J.D. Determining soil salinity and detecting saline seeps using an inductive electromagnetic soil conductivity sensor (Eng) [Determinação da salinidade do solo e identificação de nascentes salinas por meio de um sensor indutivo eletromagnético de condutividade do solo] In: *Agronomy Abstracts: 1978 Annual Meeting of the Soil Science Society of America*: 183 (USDA, SEA, Riverside, CA, USA)

   Desenvolveu-se um novo instrumento para determinar a salinidade do solo e a descoberta de nascentes salinas a partir de medições da condutividade elétrica do solo, sem sondas ou contato de terra, mediante uma técnica indutiva magnética. A condutividade é lida diretamente no instrumento e as medições podem ser feitas caminhando-se sobre o solo. Tecem-se considerações sobre o equipamento e os resultados. Examinam-se as vantagens e limitações do novo método e de métodos anteriores.

4. Gisser, M.; Pohoryles, S. Water shortage in Israel: long-run policy for the farm sector (Eng) [Escassez de água em Israel: política de longo prazo para o setor agrícola] *Water Resources*, Dec 1977, 13(6):865-872, 1 fig. 10 tab, 4 ref (University of New Mexico, Dept of Economics, Albuquerque, NM 87131, USA)

   Israel defronta uma situação de limitado volume de provisão de água e demandas crescentes. Como a agricultura utiliza uma grande parcela da água disponível, uma política potencial é reduzir as destinações de água para a agricultura, a fim de permitir o aumento de uso em outros setores. Fazem-se estimativas da perda total em rendimentos na agricultura causada pela redução das cotas atuais, empregando um modelo de programação linear.

5. Debrivna, I.Ye. (Sulfate reducing bacteria of rice irrigation systems in the Southern Ukrainian SSR) (Ukr, summary Eng) [Bactérias redutoras de sulfato em sistemas de irrigação de arroz na RSS da Ucrânia Meridional] *Mikrobiologii Jurnal*, 1977, 39(5): 627-629, 2 tab, 9 ref (Academy of Sciences of the Ukrainian SSR, Institute of Microbiology and Virology, Kiev, USSR)

   Os estudos relatados mostraram um desenvolvimento muito intenso de bactérias redutoras de sulfato no subsolo dos sistemas de irrigação de arroz caracterizados por um lençol freático alto. Sugere-se que isso seria responsável pelas quedas da produção de arroz nessas condições.

6. Koo, J.W.; Ryu, H.Y. (A study on the determination method of pumping rates in tubewells for irrigation) (Kor, summary Eng) [Um estudo sobre o método de determinação de coeficientes de bombeamento em poços tubulares para irrigação] *Journal of Korean Society of Agricultural Engineers*, Dec 1976, 18(4): 1-9, 8 fig, 4 tab, 20 ref (Seoul National University, Suweon, Republic of Korea)

   Realizaram-se ensaios de bombeamento em 12 poços tubulares com a finalidade de encontrar um método para calcular o coeficiente de bombeamento em poços tubulares para irrigação. Uma bomba centrífuga de 3", um motor de 5 hp e um entalhe em V foram empregados no ensaio, sendo medidas as profundidades, os níveis de água estática, os níveis de bombeamento e as vazões dos poços tubulares. Observou-se uma correlação negativa entre coeficiente de bombeamento e rebaixamento, e uma correlação positiva entre coeficiente de bombeamento e coeficiente de transmissibilidade. Verificou-se que uma fórmula derivada da teoria de Thiem era satisfatória para calcular os coeficientes de bombeamento de poços tubulares.

7. Shanmugarajah, K.; Atukorale, S.C. Water management at Rajangana scheme – lessons from cultivation – Yala 1976 (Eng) [Manejo hídrico no projeto Rajangana – lições do plantio – Yala 1976] *Jalavrudhi (Sri Lanka)*, Dec 1976, 1 (2): 60-65, 5 tab (Water Management Division, Irrigation Dept. Sri Lanka)

   Esta é uma descrição de como foi comprovado que os plantadores de arroz de uma certa área sempre desperdiçaram água. Durante uma seca foram convocados especialistas em hidrologia, em virtude do temor de perda da safra, e, graças à melhoria da eficiência na utilização da água, o consumo foi reduzido drasticamente, sem que houvesse redução na produção agrícola.

8. Arbarb, M.; Manbeck, D.M. Influence of lateral depth and spacing on corn yield and water use in subsurface irrigation system (Eng) [Influência da profundidade lateral e do espaçamento na produção de milho e utilização da água em sistema de irrigação subsuperficial] *Annual Meeting, ASAE, North Carolina State University, Raleigh, NC, USA, Jun 26-29, 1977, Paper No. 77-2021*, 21 p. 8 fig., 1 tab, 9 ref. Available from ASAE, POB 410, St. Joseph, MI 49085, USA (University of Nebraska, Agricultural Engineering Dept, NB, USA)

   Os objetivos desse experimento foram estudar a influência de diferentes profundidades laterais e espaçamentos na produção de milho e utilização de água, e estudar a utilização prática de um sistema de irrigação subsuperficial e o padrão de distribuição da água.

Comentários deste autor

1. A primeira frase nada acrescenta ao título. O resumo poderia ser ainda mais condensado, sem perda de sentido, como segue:

   Propõe um método que não requer agentes químicos. Estudaram-se as características hidráulicas de vários crivos, com e sem vibração, e se determinaram seus coeficientes de resistência.

2. Novamente ocorre repetição do título. Poderia ficar mais compacto assim:

   Relatam-se experimentos com várias culturas de hortaliças. Apresenta uma equação para calcular, em determinada área, o número necessário de unidades de aspersores e a duração da irrigação. Apresenta dados sobre danos ao plantio causados pelas rodas dos aspersores.
   (NB. Seria muito melhor identificar as culturas, por exemplo, 'Relatam-se experimentos com repolho, beterraba e cenoura'.)

3. Pode-se evitar repetição desnecessária e o resumo se tornaria mais conciso:

   O novo instrumento descrito funciona por meio da medição da condutividade elétrica do solo, sem sondas ou contato com a terra. Pode-se ler diretamente a condutividade e as medições feitas caminhando-se sobre o solo. Comparam-se o instrumento e seus resultados com métodos anteriores.

4. Desnecessariamente prolixo. Poderia ser reduzido a:

   Uma das formas de atenuar a escassez de água seria reduzir as cotas atribuídas à agricultura (um grande consumidor), a fim de permitir o aumento do uso em outros setores. Emprega-se um modelo de programação linear para calcular a renda agrícola que se perderia no caso de redução das cotas atuais.
   (NB. Como o título informa sobre o contexto — escassez de água em Israel — não é preciso repeti-lo no resumo. O título e o resumo se complementam; este não deve existir separado do título. Este resumo é muito prolixo: 'limitado volume de provisão de água e demandas crescentes' é um circunlóquio para dizer 'escassez de água', o que já está implícito no título.)

5. Este pode ser reduzido em quase 50%:

   Um desenvolvimento muito intenso das bactérias no subsolo de sistemas de irrigação de aqüífero alto pode ser responsável pelas quedas na produção de arroz.

6. Pode ser abreviado ainda mais:

   Empregaram-se uma bomba centrífuga de 3", um motor de 5 hp e um entalhe em V de 90 graus para medir as profundidades, níveis de água estática, níveis de bombeamento e vazões de 12 poços tubulares. O coeficiente de bombeamento correlaciona-se positivamente com o coeficiente de transmissibilidade, e negativamente com o rebaixamento. Pode-se utilizar uma fórmula derivada da teoria de Thiem para calcular os coeficientes de bombeamento.

7. Um resumo muito prolixo. Pode-se abarcar a essência do texto assim:

   Especialistas em hidrologia, convocados durante uma seca, demonstraram que a eficiência de utilização da água podia melhorar grandemente, causando uma redução drástica do consumo sem reduzir a produção de arroz.
   (NB. Várias partes do resumo original são supérfluas. A primeira frase acha-se implícita na última "mediante melhoria da eficiência na utilização da água". "Devido ao receio de perda da safra" é evidente por si mesmo e nada acrescenta ao resumo. Por outro lado, como o título é inespecífico, dever-se-ia especificar a cultura (arroz) e não 'safra' em geral. Evidentemente, não se poderia trocar 'safra' por 'arroz' sem ver o artigo original.)

8. Raro exemplo de um resumo muito ruim do *Irricab*. Não acrescenta praticamente nada à informação do título. Não seria possível melhorá-lo sem examinar o artigo original.

APÊNDICE 1

Síntese de princípios de redação de resumos*

Princípios gerais

1. Não se deve impor restrição à extensão absoluta do resumo. Deve ter a extensão que for necessária para que seja o enunciado mais direto, conciso e homogêneo possível, que inclua todas as informações positivas constantes do artigo e nenhuma informação nula. Informação nula quer dizer: 1) os elementos que se consideram sem qualquer probabilidade razoável de, direta ou indiretamente, apoiar uma decisão de trabalho; 2) os elementos que duplicam outros elementos já incluídos; e 3) os elementos que constituem conhecimento de domínio comum dos especialistas do setor.
2. Exigem-se frases curtas, bem redigidas e completas para fácil acesso à informação.
3. O resumo pode usar palavras diferentes das do artigo original [paráfrase] ou adotar, seletiva e cuidadosamente, as mesmas palavras do artigo. Quanto mais bem organizado e redigido o artigo original, maior será a dependência em relação ao último método, que é uma forma de elaboração de 'extratos'.
4. Palavras e expressões técnicas devem ser as correntes na ciência em causa.
5. Novos termos ou denominações devem ser apresentados com suas definições.
6. Somente devem ser empregados as abreviaturas e símbolos convencionais mais comuns, a fim de evitar confusão e contribuir para a legibilidade.

Princípios relativos ao conteúdo

1. O tópico introdutório deve oferecer uma indicação exata do assunto tratado e dos métodos empregados, caso isso não esteja evidente no título. Esse tópico, no entanto, será uma redundância perdulária, se o título tiver representado satisfatoriamente o conteúdo temático e o método de pesquisa.
2. Se não estiver evidente no título e/ou no tópico introdutório, o tópico seguinte deverá indicar o âmbito do artigo e a finalidade e objetivos do autor. Se o leitor do resumo estiver buscando uma informação específica, esses dois tópicos deverão indicar-lhe a probabilidade de achar a informação que lhe serve.

   De fato, esses tópicos de abertura devem ser um resumo descritivo conciso que se usa na maioria dos casos para ajudar o leitor a decidir se deve se reportar ao artigo original, mas neste caso para indicar-lhe se as informações contidas são as que busca ou se são adequadas a seu trabalho.
3. Se o artigo for de caráter experimental ou teórico, a hipótese do autor será declarada explicitamente, caso não esteja evidente nos tópicos de abertura.
4. Os métodos de pesquisa adotados devem ser identificados. Se forem empregados técnicas ou procedimentos convencionais, não é preciso descrevê-los. Se os procedimentos forem novos ou contiverem características originais aplicadas a processos conhecidos, estes aspectos serão claramente descritos. Devem ser mencionados os princípios básicos de métodos ou tecnologias novas, suas aplicações e qualidades, faixas de operação e graus de exatidão.
5. Descrever minuciosamente os métodos de coleta de dados e medidas, rotação de variáveis, método de isolamento dos dados, identificação de índices, técnicas de condensação de dados, etc. O resumidor depende do método de coleta de dados, junto com o de pesquisa, para avaliar a qualidade do trabalho do autor e a confiabilidade e validade de resultados e conclusões.
6. Os dados, sejam eles uma coleção de resultados experimentais ou argumentos teóricos, serão apresentados na medida em que, e somente na medida em que, representem integralmente todos os aspectos importantes do artigo, e sejam suficientes para conduzir logicamente às conclusões do autor. Os dados de natureza absoluta serão apresentados com detalhes suficientes que atendam ao uso que previsivelmente terão em atividades científicas futuras.

   Os dados podem ser apresentados de qualquer forma, com base no seguinte critério: adote a apresentação mais econômica possível, porém a mais lúcida. Podem-se incluir tabelas, diagramas, gráficos, desde que identificados exatamente, mas os dados assim apresentados devem se bastar, isto é, ser compreensíveis sem necessidade de recorrer ao texto do resumo.
7. Devem ser indicados os métodos qualitativos e/ou quantitativos adotados no tratamento dos dados. Não é preciso descrever técnicas convencionais e conhecidas. Variações ou aplicações especiais de técnicas conhecidas serão apresentadas se forem necessárias para representar por completo os aspectos importantes do estudo e fundamentar inteiramente as conclusões alcançadas.
8. Devem ser apresentadas as conclusões lógicas. Hipóteses e teorias serão reexaminadas se foram comprovadas ou invalidadas, aceitas ou refutadas. Neste ponto, cabe ao resumidor discriminar entre conclusões comprovadas e não-comprovadas e conclusões reais *versus* inferências. Acima de tudo, não deve apresentar conclusões que não possam ser confirmadas pelas partes anteriores do resumo. Não deve incluir proposições errôneas contidas no artigo, a não ser que sejam acompanhadas de uma advertência que de modo claro chame atenção para o erro e, se possível, para sua correção.
9. É possível incluir interpretações válidas e importantes que o autor faça sobre resultados e/ou conclusões apresentados, caso representem um avanço dos conhecimentos ao revelar novas relações ou reafirmar relações antigas.
10. Em todo o resumo, o resumidor deve exercer seu direito de esclarecer e simplificar elementos contidos no artigo.

* Síntese de princípios de redação de resumos proposta por Payne et al. (1962). Reproduzida com permissão de American Institutes for Research.

APÊNDICE 2

Análise de conteúdo modular com módulos temáticos

Citação

STOLL, A.M.; CHIANTA, M.A.; MUNROE, L.R. Flame-contact studies. *Transactions of the ASME, Series C, Journal of Heat Transfer*, vol. 86, No. 3, August 1964, p. 449-456.

Resumo

Descrevem-se aparelho e métodos de aquecimento por contato de chama, aplicados com êxito na determinação das temperaturas de destruição e características térmicas de materiais fibrosos e plásticos. Apresentam-se resultados de ensaios que confirmam a análise. Informam-se os resultados concernentes a uma fibra de poliamido, e ao efeito de isolamento de espaços de ar entre camadas de tecido.

Modelos de chapa composta foram injetados na chama de um maçarico de Meker, e se determinaram, opticamente ou mediante pares térmicos, as temperaturas da parede posterior. O fluxo de calor para a superfície foi determinado opticamente. No lado da chama da chapa composta, avaliou-se um tecido de poliamido (du Pont HT-1) de pesos variáveis por unidade de área superficial (3 onças–5 onças/jarda quadrada). O lado posterior, ou material de referência, da parede consistia num composto resinoso (pele simulada) de propriedades térmicas e ópticas conhecidas. As temperaturas de destruição do tecido HT-1 foram de 427±3 °C mediante determinação óptica e de 423±27 °C determinada por mensurações com pares térmicos. A temperatura da chama era de 1 200 °C. A queima completa ocorreu em 3–6 segundos dependendo do peso. Ao pesquisar a utilização de espaços de ar como camadas isolantes entre camadas do tecido, intervalos de 4 mm pareceram ser o ideal para o material de 3 onças/jarda quadrada. Concluiu-se que para aplicações de curta duração a altas temperaturas, materiais isolantes desse tipo seriam o ideal para proteção pessoal. Nos ensaios de validação da análise matemática foram utilizadas amostras de borracha de silicone RTV-20 muito finas (0,050–0,100 cm). Obteve-se excelente concordância entre as temperaturas de parede calculadas e as medidas (diferença percentual de 0,5 por cento); a análise adotada foi a de Griffith e Horton.

Analisa-se o emprego dessas técnicas analíticas e experimentais em relação à determinação da difusividade e condutividade térmicas de ensaios do tipo de contato de chama. Conclui-se que as técnicas proporcionaram um meio sensível e exato para determinar propriedades térmicas.

Módulos temáticos especializados

(parágrafos suplementares ao resumo básico)

Fisiologia e medicina

Descreve-se um aparelho e se desenvolvem expressões matemáticas que permitem uma análise de dano tissular, devido a exposição a chama, a partir do conhecimento das propriedades e da história de temperatura–tempo de uma camada de revestimento de produto têxtil. Isso constitui um meio relativamente simples de estudar as propriedades térmicas (inclusive difusividade e condutividade) de tecido vivo intacto sem alteração do próprio tecido.

Indústria de plásticos

HT-1, uma fibra têxtil experimental de poliamido resistente ao calor, foi exposta a contato de chama num maçarico de Meker com uma temperatura na chama de 1 200 °C. A temperatura de destruição dos tecidos de 3, 4, 5 e 6 onças/jarda quadrada foi de 427±3 °C, medida radiometricamente. A queima total ocorreu em 3–6 segundos, dependendo do peso.

Indústria da borracha

Mediu-se, por meio de um calorímetro de contato de chama, o fluxo de calor transitório através de uma montagem de duas camadas de RTV-20, uma borracha de silicone fabricada pela General Electric, reforçadas com pele simulada. Mediu-se, no interior da camada de reforço, a elevação de temperatura em três segundos em camadas de borracha de 0,95, 0,55 e 0,52 mm, o que concordou de modo excelente com os valores teóricos.

Indústrias de roupas de proteção e aeronáutica

Os experimentos descritos, sobre as temperaturas de destruição e características térmicas de tecidos submetidos ao calor por contato de chama, são da maior importância para o projeto de roupas de proteção contra queimaduras. Em particular, ajudam a explicar por que, em experiências com macacões de vôo, obtém-se significativo aumento da proteção contra queimaduras com roupas de camadas duplas em comparação com vestuário de uma única camada.

Entradas de índice

Sistemas físicos e matemáticos

CHAPAS COMPOSTAS
CHAPAS DE CAMADA ÚNICA

Transferência de calor

CONDUÇÃO TRANSIENTE ANALÍTICA
CONDUÇÃO TRANSIENTE (GRIFFITH-HORTON)
CONDUÇÃO UNIDIMENSIONAL

Meios e métodos

APARELHO EXPERIMENTAL
CALORÍMETROS DE CONTATO DE CHAMA

Outras etiquetas de assuntos
ROUPAS DE PROTEÇÃO
ROUPAS DE VÔO
QUEIMADURAS

Ambiente
TEMPERATURA: 0–1000 °F
MAÇARICO DE MEKER
CONTATO DE CHAMA

Materiais e propriedades
TECIDOS
HT-1
POLIAMIDOS
RTV-20
BORRACHA DE SILICONE
PELE
PROPRIEDADES ISOLANTES
CONDUTIVIDADE TÉRMICA
DIFUSIVIDADE TÉRMICA
PROTEÇÃO CONTRA QUEIMADURAS

Autores
STOLL, A.M.
CHIANTA, M.A.
MUNROE, L.R.S

Afiliações
Aviation Medical Acceleration Laboratory, U.S. Naval Air Development Center, Johnsville, Pennsylvania

REFERÊNCIAS

Acorn, T.L.; Walden, S.H. SMART: support management automated reasoning technology for Compaq customer service. In: *Innovative applications of artificial intelligence* 4; ed. by A.C. Scott; P. Klahr, p. 3-18. Cambridge, MA, MIT Press, 1992.

Acton, P. Indexing is *not* classifying — and vice versa. *Records Management Quarterly*, 20 (3), 1986, 10-15.

Adami, N. et al. The ToCAI description scheme for indexing and retrieval of multimedia documents. *Multimedia Tools and Applications*, 14, 2001, 153-173.

Addison, E.R. Large scale full text retrieval by concept indexing. In: *Proceedings of the Twelfth National Online Meeting*, p. 5-15. Medford, NJ, Learned Information, 1991.

Agnew, B. et al. Multi-media indexing over the Web. In: *Storage and retrieval for image and video databases V*; ed. by I.K. Sethi; R.C. Jain, p. 72-83. Bellingham, WA, International Society for Optical Engineering, 1997.

Agosti, M.; Smeaton, A.F., ed. *Information retrieval and hypertext*. Boston, Kluwer, 1996.

Agosti, M. et al. Automatic authoring and construction of hypermedia for information retrieval. *Multimedia Systems*, 3, 1995, 15-24.

Ahlswede, T. et al. Automatic construction of a phrasal thesaurus for an information retrieval system from a machine readable dictionary. RIAO 88 Conference Proceedings, v. I, p. 597-608. Paris, C.I.D., 1988.

Aitchison, J.; Cleverdon, C.W. *A report on a test of the index of metallurgical literature of Western Reserve University*. Cranfield, UK, College of Aeronautics, 1963.

Aitchison, T.M. et al. *Comparative evaluation of index languages*. London, Institution of Electrical Engineers, 1969-1970. 2 v.

Ajiferuke, I.; Chu, C.M. Quality of indexing in online databases: an alternative measure for a term discriminating index. *Information Processing & Management*, 24, 1988, 599-601.

Al-Kofahi, K. et al. Combining multiple classifiers for text categorization. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, p. 97-103. New York, Association for Computing Machinery, 2001.

Albright, J.B. *Some limits to subject retrieval from a large published index*. Doctoral thesis. Urbana-Champaign, University of Illinois, Graduate School of Library Science, 1979.

Allan, J. Knowledge management and speech recognition. *Computer*, 35(4), 2002, 60-61.

Allan, J. et al. Temporal summaries of news topics. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 10-18. New York, Association for Computing Machinery, 2001.

Alto Comissariado das Nações Unidas para os Refugiados. Refugee Documentation Centre. *A guide for abstractors*. Geneva, United Nations High Commissioner for Refugees, 1985.

Anderson, J.D. Indexing systems: extensions of the mind's organizing power. In: *Information and behavior*. V. 1; ed. by B.D. Ruben, p. 287-323. New Brunswick, N.J., Transaction Books, 1985.

Anderson, J.D.; Pérez-Carballo, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. *Information Processing & Management*, 37, 2001, 231-277.

Anderson, J.D.; Rowley, F.A. Building end-user thesauri from full-text. *Advances in Classification Research*, 2, 1992, 1-13.

Anderson, M.D. *Book indexing*. Cambridge, UK, Cambridge University Press, 1971. (Reimpresso com correções em 1979.)

Anick, P.G. Integrating natural language processing and information retrieval in a troubleshooting help desk. *IEEE Expert*, 8 (6), 1993, 9-17.

Arasu, A. et al. Searching the Web. *ACM Transactions on Internet Technology*, 1, 2001, 2-43.

Arents, H.C.; Bogaerts, W.F.L. Concept-based indexing and retrieval of hypermedia information. In: *Encyclopedia of library and information science*, v. 58, suppl. 21, p. 1-29. New York, Marcel Dekker, 1996.

Armitage, J.E.; Lynch, M.F. Some structural characteristics of articulated subject indexes. *Information Storage and Retrieval*, 4, 1968, 101-111.

Armstrong, C.J.; Keen, E.M. *Workbook for NEPHIS and KWAC*. Boston Spa, British Library, 1982. British Library Research and Development Reports Number 5710. (Microcomputer Printed Subject Indexes Teaching Package, volume 1)

Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association*, p. 17-21. Philadelphia, Hanley and Belfus, 2001.

Aronson, A.R. et al. The NLM Indexing Iniciative. *Proceedings of the 2000 Annual Symposium of the American Medical Informatics Association*, p. 17-21. Philadelphia, Hanley and Belfus, 2000.

Artandi, S. *Book indexing by computer*. Doctoral thesis. New Brunswick, NJ, Rutgers, the State University, 1963.

Aslandogan, Y.A.; Yu, C.T. Multiple evidence combination in image retrieval: Diogenes searches for people on the Web. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 88-95. New York, Association for Computing Machinery, 2000.

Austin, D. *PRECIS: a manual of concept analysis and subject indexing*. 2nd ed. London, British Library, 1984.

Austin, D.; Digger, J.A. PRECIS: the Preserved Context Index System. *Library Resources & Technical Services*, 21, 1977, 13-30.

Awre, C.; Wise, A. Portal progress. *Update*, 1(6), 2002, 46-47.

Azgaldov, E.G. A framework for description and classification of printed subject indexes. *Libri*, 19, 1969, 275-291.

Baca, M., ed. *Introduction to art image access*. Los Angeles, Getty Research Institute, 2002.

Bailin, S. et al. Application of machine learning to the organization of institutional software repositories. *Telematics and Informatics*, 10, 1993, 283-299.

Baker, S.L. Will fiction classification schemes increase use? *RQ*, 27, 1988, 366-376.

Baker, S.L.; Shepherd, G.W. Fiction classification schemes: the principles behind them and their success. *RQ*, 27, 1987, 245-251.

Bakewell, K.G.B. Reference books for indexers. *The Indexer*, 15, 1987, 131-140.

Bannan, K.J. Personalization and portals. *EContent*, 25(10), 2002, 16-21.

Bateman, J.; Teich, E. Selective information presentation in an integrated publication system: an application of genre-driven text generation. *Information Processing & Management*, 31, 1995, 753-767.

Bates, M.J. Indexing and access for digital libraries and the Internet. *Journal of the American Society for Information Science*, 49, 1998, 1185-1205.

Bates, M.J. Subject access in online catalogs: a design model. *Journal of the American Society for Information Science*, 37, 1986, 357-376.

Bates, M.J. System meets user: problems in matching subject search terms. *Information Processing and Management*, 13, 1977, 367-375.

Baxendale, P.B. Machine-made index for technical literature — an experiment. *IBM Journal of Research and Development*, 2, 1958, 354-361.

Bearman, T.C.; Kunberger, W.A. *A study of coverage overlap among fourteen major science and technology abstracting and indexing services*. Philadelphia, National Federation of Abstracting and Indexing Services, 1977.

Beghtol, C. Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, 42, 1986, 84-113.

Beghtol, C. *The classification of fiction*. Metuchen, NJ, Scarecrow Press, 1994.

Belkin, N.J. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5, 1980, 133-143.

Belkin, N.J. et al. ASK for information retrieval. *Journal of Documentation*, 38, 1982, 61-71, 145-164.

Bell, H.K. Bias in indexing and loaded language. *The Indexer*, 17, 1991a, 173-177.

Bell, H.K. Indexing fiction: a story of complexity. *The Indexer*, 17, 1991b, 251-256.

Bennett, J.L. On-line access to information: NSF as an aid to the indexer/cataloger. *American Documentation*, 20, 1969, 213-220.

Bennett, J.L. et al. *Observing and evaluating an interactive process: a pilot experiment in indexing*. San Jose, CA, IBM Research Laboratory, 1972.

Benois-Pineau, J. et al. Query by synthesized sketch in architectural database. In: *Storage and retrieval for image and video databases V*; ed. by I.K. Sethi; R.C. Jain, p. 361-367. Bellingham, Wa, International Society for Optical Engineering, 1997.

Benoit, G. Data mining. *Annual Review of Information Science and Technology*, 36, 2002, 265-310.

Berger, A.L.; Mittal, V.O. OCELOT: a system for summarizing web pages. *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 144-151. New York, Association for Computing Machinery, 2000.

Berner, E.S. et al. Performance of four computer-based diagnostic systems. *New England Journal of Medicine*, 330, 1994, 1792-1796.

Bernier, C.L.; Yerkey, A.N. *Cogent communication: overcoming reading overload*. Westport, CT, Greenwood Press, 1979.

Bernstein, L.M.; Williamson, R.E. Testing of a natural language retrieval system for a full text knowledge base. *Journal of the American Society for Information Science*, 35, 1984, 235-247.

Bertrand, A.; Cellier, J.-M. Psychological approach to indexing: effetcs of the operator's expertise upon indexing behaviour. *Journal of Information Science*, 21, 1995, 459-472.

Bertrand-Gastaldy, S. et al. Convergent theories: using a multidisciplinary approach to explain indexing results. *Proceedings of the American Society for Information Science*, 32, 1995, 56-60.

Besser, H. Image databases: the first decade, the present, and the future. In: *Digital image access & retrieval*; ed. by P.B. Heidorn; B. Sandore, p. 11-28. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Bhattacharyya, G. The effectiveness of natural language in science indexing and retrieval. *Journal of Documentation*, 30, 1974, 235-254.

Bhattacharyya, G. Elements of POPSI. In: *Indexing systems: concepts, models and techniques*; ed. by T.N. Rajan, p. 73-102. Calcutta, Indian Association of Special Libraries and Information Centres, 1981.

Biebricher, P. et al. The automatic indexing system AIR/PHYS—from research to application. In: *Readings in information retrieval*; ed. by K. Sparck Jones; P. Willett, p. 513-517. San Francisco, Morgan Kaufmann, 1997.

Bishop, A.P. Document structure and digital libraries: how researchers mobilize information in journal articles. *Information Processing & Management*, 35, 1999, 255-279.

Bishop, A.P. et al. Index quality study, part I: quantitative description of back-of-the-book indexes. In: *Indexing tradition and innovation*, p. 15-51. American Society of Indexers, 1990.

Blair, D.C. Some thoughts on the reported results of TREC. *Information Processing & Management*, 38, 2002, 445-451.

Blair, D.C.; Kimbrough, S.O. Exemplary documents: a foundation for information retrieval design. *Information Processing & Management*, 38, 2002, 363-379.

Blair, D.C.; Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28, 1985, 289-299.

Blum, T. et al. Audio databases with content-based retrieval. In: Intelligent multimedia information retrieval; ed. by M.T. Maybury, p. 113-135. Menlo Park, CA, AAAI Press, 1997b.

Blustein, J.; Staveley, M.S. Methods of generating and evaluating hypertext. *Annual Review of Information Science and Technology*, 35, 2001, 299-335.

Bodenreider, O.; Zweigenbaum, P. Identifying proper names in parallel medical terminologies. *Studies in health technology and informatics*, 77, 2000, 443-447.

Boguraev, B. et al. Summarisation miniaturisation: delivery of news to hand-helds. *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*. New Brunswick, NJ, Association for Computational Linguistics, 2001.

Bonham, M.D.; Nelson, L.L. An evaluation of four end-user systems for searching MEDLINE. *Bulletin of the Medical Library Association*, 76, 1988, 22-31.

Booth, A.; O'Rourke, A.J. The value of structured abstracts in information retrieval from MEDLINE. *Health Libraries Review*, 14, 1997, 157-166.

Borko, H. Toward a theory of indexing. *Information Processing & Management*, 13, 1977, 355-365.

Borko, H.; Bernick, M. Automatic document classification. *Journal of the Association for Computing Machinery*, 10, 1963, 151-162.

Borko, H.; Bernier, C.L. *Abstracting concepts and methods*. New York, Academic Press, 1975.

Borko, H.; Chatman, S. Criteria for acceptable abstracts: a survey of abstractors' instructions. *American Documentation*, 14, 1963, 149-160.

Borkowski, C.; Martin, J.S. Structure, effectiveness and benefits of LEXtractor, an operational computer program for automatic extraction of case summaries and dispositions from court decisions. *Journal of the American Society for Information Science*, 26, 1975, 94-102.

Borst, F. et al. TEXTINFO: a tool for automatic determination of patient clinical profiles

using text analysis. In: *Fifteenth Annual Symposium on Computer Applications in Medical Care*, p. 63-67. New York, McGraw Hill, 1992.

Bourne, C.P. *Characteristics of coverage by the* Bibliography of Agriculture *of the literature relating to agricultural research and development*. Palo Alto, CA, Information General Corporation, 1969a. PB 185 425.

Bourne, C.P. *Overlapping coverage of the* Bibliography of Agriculture *by fifteen other secondary sources*. Palo Alto, CA, Information General Corporation, 1969b. PB 185 069.

Boyce, B.R.; McLain, J.P. Entry point depth and online search using a controlled vocabulary. *Journal of the American Society for Information Science*, 40, 1989, 273-276.

Bradley, P. Indexes to works of fiction: the views of producers and users on the need for them. *The Indexer*, 16, 1989, 239-248.

Bradshaw, S.; Hammond, K. Constructing indices from citations in collections of research papers. *Proceedings of the American Society for Information Science*, 36, 1999, 741-750.

Brandow, R. et al. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31, 1995, 675-685.

Breaks, M.; Guyon, A. Edinburgh Engineering Virtual Library (EEVL). In: *The amazing Internet challenge*; ed. by A.T. Wells et al., p. 76-96. Chicago, American Library Association, 1999.

Brenner, C.W.; Mooers, C.N. A case history of a Zatocoding information retrieval system. In: *Punched cards: their applications to science and industry*. 2nd ed., ed. by R.S. Casey et al., p. 340-356. New York, Reinhold, 1958.

Brenner, E.H. et al. American Petroleum Institute's machine-aided indexing and searching project. *Science and Technology Libraries*, 5(1), 1984, 49-62.

Breton, E.J. Indexing for invention. *Journal of the American Society for Information Science*, 42, 1991, 173-177.

Breton, E.J. Why engineers don't use databases. *Bulletin of the American Society for Information Science*, 7(6), 1981, 20-23.

Brettle, A.J. et al. Comparison of bibliographic databases for information on the rehabilitation of people with severe mental illness. *Bulletin of the Medical Library Association*, 89, 2001, 353-362.

Brew, C.; Thompson, H.S. Automatic evaluation of computer generated text: a progress report on the TextEval project. In: *Proceedings of the Human Language Technology Workshop, March 8-11*, 1994, p. 108-113. San Francisco, Morgan Kaufmann, 1994.

Brittain, J.M.; Roberts, S.A. Rationalization of secondary services: measurement of coverage of primary journals and overlap between services. *Journal of the American Society for Information Science*, 31, 1980, 131-142.

Broer, J.W. Abstracts in block diagram form. *IEEE Transactions on Engineering Writing and Speech*, 14, 1971, 64-67.

Brown, E.W. et al. Toward speech as a knowledge resource. *IBM Systems Journal*, 40, 2001, 985-1001.

Brown, M.S. et al. A new comparison of the *Current Index to Journals in Education* and the *Education Index*: a deep analysis of indexing. *Journal of Academic Librarianship*, 25, 1999, 216-222.

Brown, P. et al. The democratic indexing of images. *New Review of Hypermedia and Multimedia*, 2, 1996, 107-120.

Browne, G.M. Indexing Web sites: a practical guide. *Internet Reference Services Quarterly*, 5(3), 2001, 27-41.

Bruza, P.D. et al. Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, 51, 2000, 1090-1105.

Burgin, R. The effect of indexing exhaustivity on retrieval performance. *Information Processing & Management*, 27, 1991, 623-628.

Burgin, R. The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Society for Information Science*, 46, 1995, 562-572.

Bürk, K. et al. INIS: manual for subject analysis. Vienna, International Atomic Energy Agency, 1996. IAEA-INIS-12 (Rev. 3)

Burke, F.G. The application of automated techniques in the management and control of source materials. *American Archivist*, 30, 1967, 255-278.

Burke, M. The use of repertory grids to develop a user-driven classification of a collection of digitized photographs. *Proceedings of the American Society for Information Science and Technology*, 38, 2001, 76-92.

Burnett, K. et al. A comparison of the two traditions of metadata development. *Journal of the American Society for Information Science*, 50, 1999, 1209-1217.

Busch, J.A. Building and accessing vocabulary resources for networked discovery and navigation. In: *Visualizing subject access for 21st century information resources*; ed. by P.A. Cochrane; E.H. Johnson, p. 93-105. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1998.

Buyukkokten, O. et al. Seeing the whole in parts: text summarization for web browsing on handheld devices. *Proceedings of the Tenth International Conference on the World Wide Web*, 2001. (http://www.db.stanford.edu/~orkut/papers/www10b/index.html)

Byrd, D.; Crawford, T. Problems of music information retrieval in the real world. *Information Processing & Management*, 38, 2002, 249-272.

Byrne, J.R. Relative effectiveness of titles, abstracts, and subject headings for machine retrieval from the COMPENDEX services. *Journal of the American Society for Information Science*, 26, 1975, 223-229.

Campbell, J.D. The case for creating a scholars portal to the Web. *ARL*, 211, August 2000, 1-4.

Carrick, C.; Watters, C. Automatic association of news items. *Information Processing & Management*, 33, 1997, 615-632.

Carroll, K.H. An analytical survey of virology literature reported in two announcement journals. *American Documentation*, 20, 1969, 234-237.

Casey, K.H. An analytical index to the Internet: dreams of Utopia. *College & Research Libraries*, 60, 1999, 586-595.

Cawkell, A.E. *A guide to image processing and picture management.* Brookfield, VT, Gower, 1994.

Cawkell, A.E. Picture-queries and picture databases. *Journal of Information Science*, 19, 1993, 409-423.

Chakrabarti, S. *Mining the Web: discovering knowledge from hypertext data*. San Francisco, Morgan Kaufmann, 2003.

Chang, G. et al. *Mining the World Wide Web*. Boston, Kluwer, 2001.

Charniak, E. Natural language learning. *ACM Computing Surveys*, 27, 1995, 317-319.

Chen, H. et al. Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 37(10), 1994, 56-73.

Chen, H. et al. Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, 46, 1995, 175-193.

Chen, H.-l. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, 52, 2001a, 260-273.

Chen, H.-l. An analysis of image retrieval tasks in the field of art history. *Information Processing & Management*, 37, 2001b, 701-720.

Chen, Z. Let documents talk to each other: a computer model for connection of short documents. *Journal of Documentation*, 49, 1993, 44-54.

Chen, Z. et al. Web mining for Web image retrieval. *Journal of the American Society for Information Science and Technology*, 52, 2001, 831-839.

Chiaramella, Y.; Kheirbek, A. An integrated model for hypermedia and information retrieval. In: *Information retrieval and hypertext*; ed. by M. Agosti; A.F. Smeaton, p. 139-178. Boston, Kluwer, 1996.

Chien, L.-F. et al. A spoken-access approach for Chinese text and speech information retrieval. *Journal of the American Society for Information Science*, 51, 2000, 313-323.

Choi, Y.; Rasmussen, E.M. Users' relevance criteria in image retrieval in American history. *Information Processing & Management*, 38, 2002, 695-726.

Chu, C.M.; Ajiferuke, I. Quality of indexing in library and information science databases. *Online Review*, 13, 1989, 11-35.

Chu, C.M.; O'Brien, A. Subject analysis: the critical first stage in indexing. *Journal of Information Science*, 19, 1993, 439-454.

Chu, H. Hyperlinks: how well do they represent the intellectual content of digital collections? *Proceedings of the American Society for Information Science*, 34, 1997, 361-368.

Chute, C.G.; Yang, Y. An evaluation of concept based latent semantic indexing for clinical information retrieval. *Sixteenth Annual Symposium on Computer Applications in Medical Care*, p. 639-643. New York, McGraw Hill, 1993.

Ciocca, G.; Schettini, R. A relevance feedback mechanism for content-based image retrieval. *Information Processing & Management*, 35, 1999, 605-632.

Clarke, C.L.A. Exploiting redundancy in question answering. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 358-365. New York, Association for Computing Machinery, 2001.

Clemencin, G. Querying the French Yellow Pages: natural language access to the directory. *Information Processing & Management*, 24, 1988, 633-649.

Cleveland, D.B.; Cleveland, A.D. *Introduction to indexing and abstracting.* 3rd ed. Englewood, CO, Libraries Unlimited, 2001.

Cleverdon, C.W. *A comparative evaluation of searching by controlled language and natural language in an experimental NASA data base*. Frascati, European Space Agency, Space Documentation Service, 1977.

Cleverdon, C.W. et al. *Factors determining the performance of index languages*. Cranfield, UK, College of Aeronautics, 1966. 3 v.

Cluley, H.J. Analytical Abstracts: user reaction study. *Proceedings of the Society for Analytical Chemistry*, 5, 1968, 217-221.

Coates, E.J. *Subject catalogues: headings and structure*. London, Library Association, 1960.

Coco, A. Full-text versus full-text plus editorial additions. *Legal Reference Services Quarterly*, 4 (2), 1984, 27-37.

Collison, R.L. *Abstracts and abstracting services*. Santa Barbara, CA, ABC-CLIO, 1971.

Collison, R.L. *Indexes and indexing*. 4th ed. New York, deGraaf, 1972. [Edição em português, baseada na segunda edição inglesa: *Índices e indexação*. Trad. de Antonio Agenor Briquet de Lemos. São Paulo, Polígono, 1971]

Conaway, C.W. *An experimental investigation of the influence of several index variables on index usability and a preliminary study toward a coefficient of index usability*. Doctoral thesis. New Brunswick, NJ, Rutgers University, Graduate School of Library Service, 1974.

Connolly, D.; Landeen, C. Toward a standard measure of index density. *KEYWORDS*, 9(2), 2001, 52-56.

Cook, M. *Archives and the computer*. London, Butterworths, 1980.

Cooper, W.S. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19, 1968, 30-41.

Cooper, W.S. Indexing documents by gedanken experimentation. *Journal of the American Society for Information Science*, 29, 1978, 107-119.

Cooper, W.S. Is inter-indexer consistency a hobgoblin? *American Documentation*, 20, 1969, 268-278.

Corridoni, J.M. et al. Image retrieval by color semantics with incomplete knowledge. *Journal of the American Society for Information Science*, 49, 1998, 267-282.

Corston-Oliver, S. Text compaction for display on very small screens. *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*, 2001. (http://research.microsoft.com/nlp/publications/NAACL2001)

Cosgrove, S.J.; Weimann, J.M. Expert system technology applied to item classification. Library Hi Tech, 10 (1/2), 1992, 33-40.

Cowie, J.; Lehnert, W. Information extraction. *Communications of the ACM*, 39(1), 1996, 80-91.

Crandall, M. Microsoft. In: *Linkage Inc's best practices in knowledge management and organizational learning handbook*, p. 89-123. Lexington, MA, Linkage Inc., 2000.

Craven, T.C. Abstracts produced using computer assistance. *Journal of the American Society for Information Science*, 51, 2000, 745-756.

Craven, T.C. Changes in metatag descriptions over time. *First Monday*, 6(10), 2001a (http://firstmonday.org/issues/issue6_10/craven/index.html)

Craven, T.C. A coding scheme as a basis for the production of customized abstracts. *Journal of Information Science*, 13, 1987, 51-58.

Craven, T.C. DESCRIPTION meta tags in public home and linked pages. *LIBRES: Library and Information Science Research Electronic Journal*, 11(2), 2001b (http://libres.curtin.edu.au/LIBRE11N2/craven.htm)

Craven, T.C. An experiment in the use of tools for computer-assisted abstracting. *Proceedings of the American Society for Information Science*, 33, 1996, 203-208.

Craven, T.C. NEPHIS: a nested-phrase indexing system. *Journal of the American Society for Information Science*, 28, 1977, 107-114.

Craven, T.C. Presentation of repeated phrases in a computer-assisted abstracting tool kit. *Information Processing & Management*, 37, 2001c, 221-230.

Craven, T.C. *String indexing*. Orlando, FL, Academic Press, 1986.

Craven, T.C. A thesaurus for use in a computer-aided abstracting tool kit. *Proceedings of the American Society for Information Science*, 30, 1993, 178-184.

Craven, T.C. Use of words and phrases from full text in abstracts. *Journal of Information Science*, 16, 1990, 351-358.

Cremmins, E.T. *The art of abstracting*. 2nd ed. Arlington, VA, Information Resources Press, 1996.

Croft, W.B.; Turtle, H.R. Text retrieval and inference. In: Text-based intelligent systems; ed. by P.S. Jacobs, p. 127-155. Hillsdale, NJ, Lawrence Erlbaum, 1992.

Cromp, R.F.; Dorfman, E. A spatial data handling system for retrieval of images by unrestricted regions of user interest. *Telematics and Informatics*, 9, 1992, 221-241.

Crowe, J.D. *Study of the feasibility of indexing a work's subjective viewpoint*. Doctoral thesis. Berkeley, University of California, 1986.

Cutter, C.A. *Rules for a dictionary catalog*. Washington, DC, Government Printing Office, 1876.

Dabney, D.P. The curse of Thamus: an analysis of full-text legal document retrieval. *Law Library Journal*, 78, 1986a, 5-40.

Dabney, D.P. A reply to West Publishing Company and Mead Data Central on The curse of Thamus. *Law Library Journal*, 78, 1986b, 349-350.

Dahlberg, I. On the theory of the concept. In: *Ordering systems for global information networks*; ed. by A. Neelameghan, p. 54-63. Bangalore, International Federation for Documentation, 1979.

Danilewitz, D.B.; Freiheit, F.E., IV. A knowledge-based system within a cooperative processing environment. In: *Innovative applications of artificial intelligence 4*; ed. by A.C. Scott; P. Klahr, p. 19-36. Cambridge, MA, MIT Press, 1992.

David, C. et al. Indexing as problem solving: a cognitive approach to consistency. *Proceedings of the American Society for Information Science*, 32, 1995, 49-55.

Davison, P.S.; Matthews, D.A.R. Assessment of information services. *Aslib Proceedings*, 21, 1969, 280-284.

Defense Documentation Center. *Abstracting of technical reports*. 1968. AD 667 000.

Demasco, P.W.; McCoy, K.F. Generating text from compressed input: an intelligent interface for people with severe motor impairments. *Communications of the ACM*, 35(5), 1992, 68-78.

Dempsey, L.; Heery, R. Metadata: a current view of practice and issues. *Journal of Documentation*, 54, 1998, 145-172.

De Ruiter, J. Aspects of dealing with digital information: "mature" novices on the Internet. *Library Trends*, 51, 2002, 199-209.

Deschâtelets, G. The three languages theory in information retrieval. *International Classification*, 13, 1986, 126-132.

DeZelar-Tiedman, C. Subject access to fiction: an application of the *Guidelines*. *Library Resources & Technical Services*, 40, 1996, 203-210.

Di Loreto, F. et al. A visual object-oriented query language for geographic information systems. In: *Database and expert systems applications*; ed. by N. Revell and A.M. Tjoa, p. 103-113. Berlin, Springer-Verlag, 1995. (Lecture Notes in Computer Science, Number 978).

Dimitroff, A.; Wolfram, D. Design issues in a hypertext-based information system for bibliographic retrieval. *Proceedings of the American Society for Information Science*, 30, 1993, 191-198.

Ding, W. et al. Performance of visual, verbal, and combined video surrogates. *Proceedings of the American Society for Information Science*, 36, 1999, 651-664.

Diodato, V.P. *Author indexing in mathematics*. Doctoral thesis. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1981.

Diodato, V.P. User preferences for features in back of book indexes. *Journal of the American Society for Information Science*, 45, 1994, 529-536.

Diodato, V.P.; Gandt, G. Back of book indexes and the characteristics of author and nonauthor indexing: report of an exploratory study. *Journal of the American Society for Information Science*, 41, 1991, 341-350.

Doraisamy, S.; Rüger, S.M. An approach towards a polyphonic music retrieval system. Paper presented at the Second Annual International Symposium on Music Information Retrieval, 2001. (http://ismir2001.indiana.edu/papers.html)

Doszkocs, T.E. CITE NLM: natural-language searching in an online catalog. *Information Technology and Libraries*, 2, 1983, 364-380.

Dovey, M.J. A technique for 'regular expression' style searching in polyphonic music. Paper presented at the Second Annual International Symposium on Music Information Retrieval, 2001. (http://ismir2001.indiana.edu/papers.html)

Down, N. Subject access to individual works of fiction: participating in the OCLC/LC fiction project. *Cataloging & Classification Quarterly*, 20 (2), 1995, 61-69.

Downie, S.; Nelson, M. Evaluation of a simple and effective music information retrieval method. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 73-80. New York, Association for Computing Machinery, 2000.

Doyle, L.B. Semantic road maps for literature searchers. *Journal of the Association for Computing Machinery*, 8, 1961, 553-578.

Drage, J.F. User preferences in technical indexes. *The Indexer*, 6, 1969, 151-155.

Driscoll, J.R. et al. The operation and performance of an artificially intelligent keywording system. *Information Processing & Management*, 27, 1991, 43-54.

Dronberger, G.B.; Kowitz, G.T. Abstract readability as a factor in information systems. *Journal of the American Society for Information Science*, 26, 1975, 108-111.

Drott, M.C. Indexing aids at corporate websites: the use of robots.txt and META tags. *Information Processing & Management*, 38, 2002, 209-210.

Dubois, C.P.R. Free text vs. controlled vocabulary: a reassessment. *Online Review*, 11, 1987, 243-253.

Dumais, S.T. Latent semantic indexing (LSI): TREC-3 report. In: *Overview of the Third Text Retrieval Conference* (TREC-3); ed. by D.K. Harman, p. 219-230. Gaithersburg, MD, National Institute of Standards and Technology, 1995. NIST Special Publication 500-225.

Dutta, S.; Sinha, P.K. Pragmatic approach to subject indexing: a new concept. *Journal of the American Society for Information Science*, 35, 1984, 325-331.

Dym, E.D. Relevance predictability: I. Investigation, background and procedures. In: *Electronic handling of information: testing and evaluation*; ed. by A. Kent et al., p. 175-185. Washington, DC, Thompson Book Co., 1967.

Earl, L.L. Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6, 1970, 313-334.

Eastman, C.M. 30,000 hits may be better than 300: precision anomalies in Internet searches. *Journal of the American Society for Information Science and Technology*, 53, 2002, 879-882.

Ebinuma, Y. et al. Promotion of keyword assignment to scientific literature by contributors. *International Forum on Information and Documentation*, 8(3), 1983, 16-20.

Eco, U. *The role of the reader: explorations in the semiotics of texts*. Bloomington, Indiana University Press, 1979. [Antologia de ensaios selecionados dos livros *Obra aberta*, *Apocalípticos e integrados*, *As formas do conteúdo*, *Lector in fabula*, *O super-homem de massa*.]

Edmundson, H.P. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16, 1969, 264-289.

Edmundson, H.P. et al. *Final report on the study for automatic abstracting*. Canoga Park, CA, Thompson Ramo Wooldridge, 1961. PB 166 532.

Edwards, T. A comparative analysis of the major abstracting and indexing services for library and information science. *Unesco Bulletin for Libraries*, 30, 1976, 18-25.

Elchesen, D.R. Cost effectiveness comparison of manual and on-line retrospective bibliographic searching. *Journal of the American Society for Information Science*, 29, 1978, 56-66.

Elhadad, N.; McKeown, K. Towards generating patient specific summaries of medical articles. Presentation at the NAACL 2001 Workshop on Automatic Summarization.

Ellis, D. et al. In search of the unknown user: indexing, hypertext and the World Wide Web. *Journal of Documentation*, 54, 1998, 28-47.

Ellis, D. et al. On the creation of hypertext links in full-text documents: measurement of inter-linker consistency. *Journal of Documentation*, 50, 1994, 67-98.

Ellis, D. et al. On the creation of hypertext links in full-text documents: measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47, 1996, 287-300.

Elrod, J.M. Classification of Internet resources: an AUTOCAT discussion. *Cataloging & Classification Quarterly*, 29(4), 2000, 19-38.

Endres-Niggemeyer, B. A naturalistic model of abstracting. In: *Advances in Knowledge Organization*, 4, 1994, 181-187.

Endres-Niggemeyer, B. *Summarizing information*. Berlin, Springer-Verlag, 1998.

Enser, P.G.B. Pictorial information retrieval. *Journal of Documentation*, 51, 1995, 126-170.

Enser, P.G.B. Visual information retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science*, 26, 2000, 199-210.

ERIC processing manual. Section 7: indexing. Washington, DC, U.S. Department of Education, Educational Resources Information Center, 1980.

Etzioni, O. The World-Wide Web: quagmire or gold mine? *Communications of the ACM*, 39(11), 1996, 65-68.

Fairthorne, R.A. Automatic retrieval of recorded information. *Computer Journal*, 1(1), 1958, 36-41.

Falk, J.D.; Baser, K.H. ABC-Spindex: a subject profile, rotated string indexing system. *Proceedings of the American Society for Information Science*, 17, 1980, 152-154.

Farradane, J. A comparison of some computer-produced permuted alphabetical subject indexes. *International Classification*, 4, 1977, 94-101.

Farradane, J. Concept organization for information retrieval. *Information Storage and Retrieval*, 3, 1967, 297-314.

Farradane, J. Relational indexing. *Journal of Information Science*, 1, 1979, 267-276; 1, 1980, 313-324.

Farradane, J.; Yates-Mercer, P.A. Retrieval characteristics of the index to *Metals Abstracts*. *Journal of Documentation*, 29, 1973, 295-314.

Fayyad, U.; Uthurusamy, R. Evolving data mining into solutions for insights. *Communications of the ACM*, 45(8), 2002, 28-31.

Feder, J.D.; Hobbs, E.T. Speech recognition and full-text retrieval: interface and integration. *Proceedings of the Sixteenth National Online Meeting*, p. 97-104. Medford, NJ, Learned Information, 1995.

Fedosyuk, M.Yu. Linguistic criteria for differentiating informative and indicative abstracts. *Automatic Documentation and Mathematical Linguistics*, 12(3), 1978, 98-110. [Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 2, 12 (9), 1978, p. 11-17.]

Feinberg, H., ed. *Indexing specialized formats and subjects*. Metuchen, NJ, Scarecrow Press, 1983.

Feiten, B.; Günzel, S. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3), 1994, 53-65.

Fidel, R. Individual variability in online search behavior. *Proceedings of the American Society for Information Science*, 22, 1985, 69-72.

Fidel, R. User-centered indexing. *Journal of the American Society for Information Science*, 45, 1994, 572-576.

Fidel, R. Who needs controlled vocabulary? *Special Libraries*, 83, 1992, 1-9.

Fidel, R. Writing abstracts for free-text searching. *Journal of Documentation*, 42, 1986, 11-21.

Fleuret, F.; Geman, D. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41, 2001, 85-107.

Flickner, M. et al. Query by image and video content: the QBIC system. Computer, 28(9), 1995, 23-32.

Floridi, L. Brave.Net.World: the Internet as a disinformation superhighway? *Electronic Library*, 14, 1996, 509-514.

Flynn, M.K. Take a letter, computer: speech recognition is coming of age. *PC Magazine*, 12(13), 1993, 29.

Forrester, M.A. Hypermedia and indexing: identifying appropriate models from user studies. In: *Online Information 93*, p. 313-324. Medford, NJ, Learned Information, 1993.

Forsyth, D.A. et al. Finding pictures of objects in large collections of images. In: *Digital image access & retrieval*; ed. by P.B. Heidorn and B. Sandore, p. 118-139. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Fowler, R.H. et al. Visualizing and browsing WWW semantic content. In: *Proceedings of the First Annual Conference on Emerging Technologies and Applications in Communications*, p. 110-113. Los Alamitos, CA, IEEE Computer Society Press, 1996.

Fox, E.A. et al. Building a large thesaurus for information retrieval. *Proceedings of the Second Conference on Applied Natural Language Processing*, p. 101-108. Morristown, NJ, Association for Computational Linguistics, 1988.

Freitas, A.A. *Data mining and knowledge discovery with evolutionary algorithms*. Berlin, Springer, 2002.

Fridman, E.P.; Popova, V.N. Otrajenie mirovoi literaturi po eksperimental'noi primatologii v *Referativnikh Jurnalakh SSSR. Nauchno-Tekhnicheskaia Informatsiia*, Seriia 1, No. 2, 1972, 34-36.

Fried, C.; Prevel, J.J. *Effects of indexing aids on indexing performance*. Bethesda, MD, General Electric Co., 1966. RADC-TR-66-525.

Friis, T. Assisted INdexing (CAIN). *IAALD Quarterly Bulletin*, 37, 1992, 35-37.

Froom, P.; Froom, J. Deficiencies in structured medical abstracts. *Journal of Clinical Epidemiology*, 46, 1993a, 591-594.

Froom, P.; Froom, J. Response to commentary by R.B. Haynes on 'Deficiencies in structured medical abstracts'.*Journal of Clinical Epidemiology*, 46, 1993b, 599.

Frost, C. The role of mental models in a multimodal image search. *Proceedings of the American Society for Information Science and Technology*, 38, 2001, 52-57.

Fugmann, R. The five-axiom theory of indexing and information supply. *Journal of the American Society for Information Science*, 36, 1985, 116-129.

Fugmann, R. Review of second edition of *Vocabulary control for information retrieval* by F.W. Lancaster. *International Classification*, 14, 1987, 164-166.

Fugmann, R. Toward a theory of information supply and indexing. *International Classification*, 6, 1979, 3-15.

Fuhr, N. Models for retrieval with probabilistic indexing. *Information Processing & Management*, 25, 1989, 55-72.

Fum, D. et al. Forward and backward reasoning in automatic abstracting. In: *COLING 82, Proceedings of the Ninth International Conference on Computational Linguistics*; ed. by J. Horecky, p. 83-88. Amsterdam, North Holland Publishing, 1982.

Funk, M.E. et al. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71, 1983, 176-183.

Gaizauskas, R.; Wilks, Y. Information extraction: beyond document retrieval. *Journal of Documentation*, 54, 1998, 70-105.

Gao, Y.J. et al. Fuzzy multilinkage thesaurus builder in multimedia information systems. In: *Proceedings of Third International Conference on Document Analysis and Recognition*. Volume 1, p. 142-145. Los Alamitos, CA, IEEE Computer Society Press, 1995.

Gardiner, D. et al. TREC-3: experience with conceptual relations in information retrieval. In: *Overview of the Third Text Retrieval Conference* (TREC-3); ed. by D.K. Harman, p. 333-352. Gaithersburg, MD, National Institute of Standards and Technology, 1995. NIST Special Publication 500-225.

Gauch, J.M. et al. Real time video scene detection and classification. *Information Processing & Management*, 35, 1999, 381-400.

Gauvain, J.-L. et al. Audio partitioning and transcription for broadcast data indexation. *Multimedia Tools and Applications*, 14, 2001, 187-200.

Gee, F.R. TIPSTER Phase III accomplishments. In: *Proceedings of the TIPSTER Text Program, Phase III*, p. 7-13. San Francisco, Morgan Kaufmann, 1999.

Geisler, G. Interface concepts for the Open Video Project. *Proceedings of the American Society for Information Science and Technology*, 38, 2001, 58-75.

Gilchrist, A. Documentation of documentation: a survey of leading abstracts services in documentation and an identification of key journals. *Aslib Proceedings*, 18, 1966, 62-80.

Girgensohn, A. et al. Keyframe-based user interfaces for digital video. *Computer*, 34(9), 2001, 61-67.

Godby, C.J. Two techniques for the identification of phrases in full text. *Journal of Library Administration*, 34, 2001, 57-65.

Godby, C.J.; Reighart, R. Terminology identification in a collection of Web resources. In: *CORC: new tools and possibilities for cooperative electronic resource description*; ed. by K. Calhoun and J.J. Riemer, p. 49-65. Binghampton, NY, Haworth Press, 2001a.

Godby, C.J.; Reighart, R. The WordSmith indexing system. *Journal of Library Administration*, 34, 2001b, 375-384.

Goldstein, J. et al. Multi-document summarization by sentence extraction. *Proceedings of the ANLP 2000 Workshop on Automatic Summarization*, p. 40-48. New Brunswick, NJ, Association for Computational Linguistics, 2000.

Gong, Y.; Liu, X. Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 19-25. New York, Association for Computing Machinery, 2001.

Goode, D.J. et al. Comparative analysis of *Epilepsy Abstracts* and a MEDLARS bibliography. *Bulletin of the Medical Library Association*, 58, 1970, 44-50.

Goodrum, A.A. Multidimensional scaling of video surrogates. *Journal of the American Society for Information Science and Technology*, 52, 2001, 174-182.

Goodrum, A.A.; Spink, A. Visual information seeking: a study of image queries on the World Wide Web. *Proceeedings of the American Society for Information Science*, 36, 1999, 665-674.

Goodrum, A.A. et al. An open source agenda for research linking text and image content features. *Journal of the American Society for Information Science and Technology*, 52, 2001, 948-953.

Gordon, M.D.; Dumais, S. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, 49, 1998, 674-685.

Gowtham, M.S.; Kamat, S.K. An expert system as a tool to classification. *Library Science with a Slant to Documentation and Information Studies*, 32(2), 1995, 57-63.

Green, A. Keeping up with the times: evaluating currency of indexing, language coverage and subject area coverage in three music periodical index databases. *Music Reference Services Quarterly*, 8(1), 2001, 53-68.

Green, B.F. et al. BASEBALL: an automatic question-answerer. In: *Computers and thought*; ed. by E. Feigenbaum and J. Feldman, p. 207-216. New York, McGraw Hill, 1963.

Green, E.-L.; Klasén, L. Indexing and information retrieval of moving images — experiences from a large television information database. In: *Online Information 93*, p. 129-136. Medford, NJ, Learned Information, 1993.

Green, R. The role of relational structures in indexing for the humanities. *Knowledge Organization*, 24, 1997, 72-83.

Greenberg, J. Metadata generation. *Bulletin of the American Society for Information Science and Technology*, 29(2), 2003, 16-19.

Greenberg, J. A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science and Technology*, 52, 2001, 917-924.

Greisdorf, H.; O'Connor, B.C. Modelling what users see when they look at images: a cognitive viewpoint. *Journal of Documentation*, 58, 2002, 6-29.

Grimson, W.E.L.; Mundy, J.L. Computer vision applications. *Communications of the ACM*, 37(3), 1994, 45-51.

Grishman, R. Whither written language evaluation? In: *Proceedings of the Human Language Technology Workshop, March 8-11, 1994*, p. 120-125. San Francisco, Morgan Kaufmann, 1994.

Guard, A. An antidote for browsing: subject headings for fiction. *Technicalities*, 11(12), 1991, 10-14.

Gudivada, V.N.; Raghavan, V.V. Content-based image retrieval systems. *Computer*, 28(9), 1995, 18-22.

Gudivada, V.N.; Raghavan, V.V. Modeling and retrieving images by content. *Information Processing & Management*, 33, 1997, 427-452.

Gudivada, V.N. et al. A unified approach to data modeling and retrieval for a class of image database applications. In: *Multimedia database systems*; ed. by V.S. Subrahmanian and S. Jajodia, p. 37-78. Berlin, Springer-Verlag, 1996.

Guenther, R.; McCallum, S. New metadata standards for digital resources: MODS and METS. *Bulletin of the American Society for Information Science and Technology*, 29(2), 2003, 16-19.

Guglielmo, E.J.; Rowe, N.C. Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems*, 14, 1996, 237-267.

Guidelines for abstracts. Bethesda, MD, National Information Standards Organization, 1997. ANSI/NISO Z39.14-1997. Reeditada em 2002.

Guidelines for indexes and related information retrieval devices (by James D. Anderson). Bethesda, MD, National Information Standards Organization, 1997. NISO-TR02-1997.

Guidelines on subject access to individual works of fiction, drama, etc. 2nd ed. Chicago, American Library Association, 2000.

Gupta, A.; Jain, R. Visual information retrieval. *Communications of the ACM*, 40(5), 1997, 71-79.

Guthrie, L. et al. Document classification and routing: a probabilistic approach. In: *Natural language information retrieval*; ed. by T. Strzalkowski, p. 289-310. Boston, Kluwer, 1999.

Haas, S.W. Natural language processing: toward large-scale, robust systems. *Annual Review of Information Science and Technology*, 31, 1996, 83-119.

Hafed, Z.M.; Levine, M.D. Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43, 2001, 167-188.

Hagerty, K. *Abstracts as a basis for relevance judgement*. Chicago, University of Chicago, Graduate Library School, 1967. Working paper no. 380-5.

Hahn, U.; Mani, I. The challenges of automatic summarization. *Computer*, 33(11), 2000, 29-36.

Hahn, U.; Reimer, U. Heuristic text parsing in 'TOPIC': methodological issues in a knowledge-based text condensation system. In: *Representation and exchange of knowledge as a basis of information processes*; ed. by H.J. Dietschmann, p. 143-163. Amsterdam, North-Holland, 1984.

Hall, A.M. *Case studies of the use of subject indexes*. London, Institution of Electrical Engineers, 1972a.

Hall, A.M. *User preferences in printed indexes*. London, Institution of Electrical Engineers, 1972b.

Han, J.; Chang, K.C.-C. Data mining for Web intelligence. *Computer*, 35(11), 2002, 64-70.

Hanson, C.W.; Janes, M. Coverage by abstracting journals of conference papers. *Journal of Documentation*, 17, 1961, 143-149.

Harman, D. The TREC conferences. In: *Readings in information retrieval*; ed. by K. Sparck Jones and P. Willett, p. 247-256. San Francisco, Morgan Kaufmann, 1997.

Harpring, P. The language of images: enhancing access to images by applying metadata schemas and structured vocabularies. In: *Introduction to art image access*; ed. by M. Baca, p. 20-39. Los Angeles, Getty Research Institute, 2002.

Harris, D. et al. *The testing of inter-indexer consistency at various indexing depths*. Chicago, University of Chicago, Graduate Library School, 1966. Working paper no. 380-2.

Hart, P.E.; Graham, J. Query-free information retrieval. *IEEE Expert*, 12(5), 1997, 32-37.

Harter, S.P. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43, 1992, 602-615.

Hartley, J. Are structured abstracts more or less accurate than traditional ones? *Journal of Information Science*, 26, 2000a, 273-277.
Hartley, J. Clarifying the abstracts of systematic literature reviews. *Bulletin of the Medical Library Association*, 88, 2000b, 332-337.
Hartley, J. Do structured abstracts take more space? And does it matter? *Journal of Information Science*, 28, 2002, 417-422.
Hartley, J. Is it appropriate to use structured abstracts in non-medical science journals? *Journal of Information Science*, 24, 1998, 359-364.
Hartley, J. Three ways to improve the clarity of journal abstracts. *British Journal of Educational Psychology*, 64, 1994, 331-343.
Hartley, J. Typographic settings for structured abstracts. *Journal of Technical Writing and Communication*, 30, 2000c, 355-365.
Hartley, J.; Benjamin, M. An evaluation of structured abstracts in journals published by the British Psychological Society. *British Journal of Educational Psychology*, 68, 1998, 443-456.
Hartley, J.; Sydes, M. Which layout do you prefer? An analysis of readers' preferences for different typographic layouts of structured abstracts. *Journal of Information Science*, 22, 1996, 27-37.
Hartley, J. et al. Obtaining information accurately and quickly: are structured abstracts more efficient? *Journal of Information Science*, 22, 1996, 349-356.
Hastings, S.K. An exploratory study of intellectual access to digitized art images. *Proceedings of the Sixteenth National Online Meeting*, p. 177-185. Medford, NJ, Learned Information, 1995a.
Hastings, S.K. Index access points in a study of intellectual access to digitized art images. In: *Multimedia computing and museums*; ed. by D. Bearman, p. 299-309. Pittsburgh, PA, Archives and Museum Informatics, 1995b.
Hastings, S.K. Query categories in a study of intellectual access to digitized art images. *Proceedings of the American Society for Information Science*, 32, 1995c, 3-8.
Haug, P.; Beesley, D. Automated selection of clinical data to support radiographic interpretation. In: *Fifteenth Annual Symposium on Computer Applications in Medical Care*, p. 593-597. New York, McGraw Hill, 1992.
Hayes, P.J. Intelligent high-volume text processing using shallow, domain-specific techniques. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 227-241. Hillsdale, NJ, Lawrence Erlbaum, 1992a.
Hayes, P.J.; Weinstein, S.P. Construe-TIS: a system for content-based indexing of a database of news stories. In: Innovative applications of artificial intelligence 2; ed. by A. Rappaport and R. Smith, p. 51-64. Cambridge, MA, MIT Press, 1991.
Hayes, S. Enhanced catalog access to fiction: a preliminary study. *Library Resources & Technical Services*, 36, 1992b, 441-459.
Haynes, R.B. More informative abstracts: current status and evaluation. *Journal of Clinical Epidemiology*, 46, 1993, 595-597.
Haynes, R.B. et al. More informative abstracts revisited. *Annals of Internal Medicine*, 113, 1990, 69-76.
Haynes, R.B. et al. Online access to MEDLINE in clinical settings: a study of use and usefulness. *Annals of Internal Medicine*, 112, 1990, 78-84.
Hearst, M.A. The use of categories and clusters for organizing retrieval results. In: *Natural language information retrieval*; ed. by T. Strzalkowski, p. 333-374. Boston, MA, Kluwer, 1999.

Heidorn, P.B. The identification of index terms in natural language object descriptions. *Proceedings of the American Society for Information Science*, 36, 1999, 472-481.
Heller, J. *On logical data organization, card catalogs, and the GRIPHOS management information system*. Rochester, NY, Margaret Woodbury Strong Museum, 1974. Museum Data Bank Research Report Number 3.
Henzler, R.G. Free or controlled vocabularies: some statistical user-oriented evaluations of biomedical information systems. *International Classification*, 5, 1978, 21-26.
Herner, S. Subject slanting in scientific abstracting publications. In: *International conference on scientific information, Washington, DC, Proceedings*. Volume 1, p. 407-427. Washington, DC, National Academy of Sciences, 1959.
Hersey, D.F. et al. Free text word retrieval and scientist indexing: performance profiles and costs. *Journal of Documentation*, 27, 1971, 167-183.
Hersh, W.R.; Hickam, D.H. A comparative analysis of retrieval effectiveness for three methods of indexing AIDS-related abstracts. *Proceedings of the American Society for Information Science*, 28, 1991, 211-225.
Hersh, W.R.; Hickam, D.H. An evaluation of interactive Boolean and natural language searching with an online medical textbook. Journal of the American Society for Information Science, 46, 1995a, 478-489.
Hersh, W.R.; Hickam, D.H. Information retrieval in medicine: the SAPHIRE experience. *Journal of the American Society for Information Science*, 46, 1995b, 743-747.
Hersh, W.R. et al. Words, concepts, or both? Optimal indexing units for automated information retrieval. *Sixteenth Annual Symposium on Computer Applications in Medical Care*, p. 644-648. New York, NY, McGraw Hill, 1993.
Hert, C.A. et al. A usability assessment of online indexing structures in the networked environment. *Journal of the American Society for Information Science*, 51, 2000, 971-988.
Hickey, T.B.; Vizine-Goetz, D. The role of classification in CORC. *Journal of Library Administration*, 34, 2001, 421-430.
Hidderley, R.; Rafferty, P. Democratic indexing: an approach to the retrieval of fiction. *Information Services & Use*, 17, 1997, 101-109.
Hill, L.L. Collection of metadata solutions for digital library applications. *Journal of the American Society for Information Science*, 50, 1999, 1169-1181.
Hinman, H.; Leita, C. Librarians Index to the Internet (LII). In: *The amazing Internet challenge*; ed. by A.T. Wells et al., p. 144-160. Chicago, IL, American Library Association, 1999.
Hjørland, B. Relevance research. *Journal of the American Society for Information Science and Technology*, 51, 2000, 209-211.
Hjørland, B. Toward a theory of aboutness, subject, topicality, theme, domain, field, content... and relevance. *Journal of the American Society for Information Science and Technology*, 52, 2001, 774-778.
Hjørland, B.; Nielsen, L.K. Subject access points in electronic retrieval. *Annual Review of Information Science and Technology*, 35, 2001, 249-298.
Hlava, M.M.K. Machine-aided indexing (MAI) in a multilingual environment. In: *Online Information 92*, p. 297-300. Medford, NJ, Learned Information, 1992.
Hmeidi, I. et al. Design and implementation of automatic indexing for information retrieval with Arabic documents. *Journal of the American Society for Information Science*, 48, 1997, 867-881.

Hobbs, J.R.; Israel, D. Principles of template design. In: *Proceedings of the Human Language Technology Workshop, March 8-11, 1994*, p. 177-181. San Francisco, CA, Morgan Kaufmann, 1994.

Hobbs, J.R. et al. Robust processing of real-world natural-language texts. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 13-33. Hillsdale, NJ, Lawrence Erlbaum, 1992.

Hock, R.E. *The extreme searcher's guide to web search engines*. 2nd ed. Medford, NJ, Information Today, 2001.

Hock, R.E. Sizing up HotBot: evaluating one Web search engine's capabilities. *Online*, 21(6), 1997, 24-33.

Hodges, P.R. Keyword in title indexes: effectiveness of retrieval in computer searches. *Special Libraries*, 74, 1983, 56-60.

Hogan, M. et al. The visual thesaurus in a hypermedia environment. In: *Hypermedia & interactivity in museums*; ed. by D. Bearman, p. 202-221. Pittsburgh, PA, Archives and Museum Informatics, 1991.

Holm, B.E.; Rasmussen, L.E. Development of a technical thesaurus. *American Documentation*, 12, 1961, 184-190.

Holmes, N. The KWIC and the dead: a lesson in computing history. *Computer*, 34(1), 2001, 142-144.

Holst, W. Problemer ved strukturering og bruk av den polytekniske tesaurus. *Tidskrift for Dokumentation*, 22, 1966, 69-74.

Holt, B.; Hartwick, L. 'Quick, who painted fish?': searching a picture database with the QBIC project at UC Davis. *Information Services & Use*, 14, 1994, 79-90.

Holt, B. et al. The QBIC project in the Department of Art and Art History at UC Davis. *Proceedings of the American Society for Information Science*, 34, 1997, 189-195.

Holt, G.E. On becoming essential: an agenda for quality in twenty-first century public libraries. *Library Trends*, 44, 1995, 545-571.

Hooper, R.S. Evaluation and analysis of indexing systems. In: *The Second Institute on Technical Literature Indexing*, Session 1. Washington, DC, American University, Center for Technology and Administration, 1966.

Hooper, R.S. *Indexer consistency tests — origin, measurements, results and utilization*. Bethesda, MD, IBM, 1965.

Horký, J. Shoda mezi zpracovateli pri vyberu klicovych slov z odbornych textu. [Concordância na seleção de palavras-chave de textos especializados.] *Ceskoslovenská Informatika*, 25, 1983, 275-278.

Horty, J.F. Experience with the application of electronic data processing systems in general law. *Modern Uses of Logic in Law*, 60D, 1960, 158-168.

Horty, J.F. Legal research using electronic techniques. In: *Literature of the law — techniques of access*, p. 56-68. South Hackensack, NJ, F.B. Rothman & Co., 1962.

Hourihane, C. It begins with the cataloger: subject access to images and the cataloguer's perspective. In: *Introduction to art image access*; ed. by M. Baca, p. 40-66. Los Angeles, Getty Research Institute, 2002.

Hovy, E. Using an ontology to simplify data access. *Communications of the ACM*, 46(1), 2003, 47-49.

Huang, T. et al. Multimedia Analysis and Retrieval System (MARS) project. In: *Digital image access & retrieval*; ed. by P.B. Heidorn and B. Sandore, p. 100-117. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Hui, S.C.; Goh, A. Incorporating abstract generation into an online retrieval interface for a library newspaper cutting system. *Aslib Proceedings*, 48, 1996, 259-265.

Humphrey, S.M. Automated indexing. *Bulletin of the American Society of Indexers*, 8, 2000, 157-159.

Humphrey, S.M. Automatic indexing of documents from journal descriptors: a preliminary investigation. *Journal of the American Society for Information Science*, 50, 1999, 661-674.

Humphrey, S.M. Interactive knowledge-based systems for improved subject analysis and retrieval. In: *Artificial intelligence and expert systems: will they change the library?*; ed. by F.W. Lancaster and L.C. Smith, p. 81-117. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1992.

Humphrey, S.M. Comunicação pessoal por correio eletrônico, 20 de novembro de 1995.

Humphrey, S.M. et al. Automatic indexing by discipline and high-level categories. A ser publicado em *Advances in Classification Research*, 11, 2003 (no prelo).

Humphreys, K. et al. Bioinformatics applications of information extraction from scientific journal articles. *Journal of Information Science*, 26, 2000, 75-85.

Hurt, C.; Potter, W.G. CORC and the future of libraries. In: *CORC new tools and possibilities for cooperative electronic resource description*; ed. by K. Calhoun and J.J. Riemer, p. 17-27. Binghampton, NY, Haworth Press, 2001.

Hutchins, W.J. The concept of 'aboutness' in subject indexing. *Aslib Proceedings*, 30, 1978, 172-181.

Intner, S.S. Censorship in indexing. *The Indexer*, 14, 1984, 105-108.

Introna, L.; Nissenbaum, H. Defining the Web: the politics of search engines. Computer, 33(1), 2000, 54-62.

Irving, H.B. Computer-assisted indexing training and electronic text conversion at NAL. *Knowledge Organization*, 24, 1997, 4-7.

Iyengar, S.S. Visual based retrieval systems and Web mining. *Journal of the American Society for Information Science and Technology*, 52, 2001, 828-875.

Iyer, H.; Giguere, M. Towards designing an expert system to map mathematics classificatory structures. *Knowledge Organization*, 22, 1995, 141-147.

Jackson, M.E. The advent of portals. *Library Journal*, 127(15), 2002, 36-39.

Jacobs, P.S. Introduction: text power and intelligent systems. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 1-8, Hillsdale, NJ, Lawrence Erlbaum, 1992a.

Jacobs, P.S. Joining statistics with NLP for text categorization. In: *Proceedings of the Third Conference on Applied Natural Language Processing*, p. 178-185. San Francisco, Morgan Kaufmann, 1992b.

Jacobs, P.S., ed. *Text-based intelligent systems: current research and practice in information extraction and retrieval*. Hillsdale, NJ, Lawrence Erlbaum, 1992c.

Jacobs, P.S.; Rau, L.F. Innovations in text interpretation. In: *Natural language processing*; ed. by F.C.N. Pereira and B.J. Grosz, p. 143-191. Cambridge, MA, MIT Press, 1994.

Jacobs, P.S.; Rau, L.F. SCISOR: extracting information from online news. *Communications of the ACM*, 33(11), 1990, 88-97.

Jacoby, J.; Slamecka, V. *Indexer consistency under minimal conditions*. Bethesda, MD, Documentation Inc., 1962. RADC-TDR-62-426.

Jacsó, P. Document-summarization software. *Information Today*, 19(2), 2002, 22-23.

Jagadish, H.V. Indexing for retrieval by similarity. In: *Multimedia database systems*; ed. by V.S. Subrahmanian and S. Jajodia, p. 165-184. Berlin, Springer-Verlag, 1996.

Jahoda, G; Stursa, M.L. A comparison of a kewyord from title index with a single access point per document alphabetic subject index. *American Documentation*, 20, 1969, 377-380.

Jain, R. Visual information retrieval in digital libraries. In: *Digital image access & retrieval*; ed. by P.B. Heidorn and B. Sandore, p. 68-85. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Janes, J.W. Relevance judgments and the incremental presentation of document representations. *Information Processing & Management*, 27, 1991, 629-646.

Jansen, B.J.; Pooch, U. A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52, 2001, 235-246.

Johnson, F.C. et al. The application of linguistic processing to automatic abstract generation. In: *Readings in information retrieval*; ed. by K. Sparck Jones and P. Willett, p. 538-551. San Francisco, CA, Morgan Kaufmann, 1997.

Jonak, Z. Problemy informacni analyzy pri Popisu Beletristckeho Dila. [Problemas de análise da informação na descrição de uma obra de ficção.] *Kniznice a Vedecke Informacie*, 10(1), 1978, 16-21.

Jones, E.K.; Roydhouse, A. Intelligent retrieval of archived meteorological data. *IEEE Expert*, 10(6), 1995, 50-57.

Jones, K.P. Towards a theory of indexing. *Journal of Documentation*, 32, 1976, 118-125.

Jones, K.P.; Bell, C.L.M. Artificial intelligence program for indexing automatically (AIPIA). In: *Online Information 92*, p. 187-196. Medford, NJ, Learned Information, 1992.

Jones, S.; Paynter, G.W. Automatic extraction of document keyphrases for use in digital libraries. *Journal of the American Society for Information Science and Technology*, 53, 2002, 653-677.

Jonker, F. *Indexing theory, indexing methods and search devices*. New York, NY, Scarecrow Press, 1964.

Jörgensen, C. Indexing images: testing an image description template. *Proceedings of the American Society for Information Science*, 33, 1996, 209-213.

Jörgensen, C. Introduction and overview. *Journal of the American Society for Information Science and Technology*, 52, 2001, 906-910.

Kaiser, J.O. *Systematic indexing*. London, Pitman, 1911.

Karasev, S.A. Abstracting scientific and technical literature: elements of a theory. *Automatic Documentation and Mathematical Linguistics*, 12(4), 1978, 1-7. [Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 2, 12(10), 1978, p. 1-4.]

Kassirer, J.P. A report card on computer-assisted diagnosis — the grade: C. *New England Journal of Medicine*, 330, 1994, 1824-1825.

Katzer, J. et al. A study of the overlap among document representations. *Information Technology: Research and Development*, 1, 1982, 261-274.

Keen, E.M. On the generation and searching of entries in printed subject indexes. *Journal of Documentation*, 33, 1977a, 15-45.

Keen, E.M. On the processing of printed subject index entries during searching. *Journal of Documentation*, 33, 1977b, 266-276.

Keen, E.M. Query term weighting schemes for effective ranked output retrieval. *Online Information 91*, p. 135-142. Medford, NJ, Learned Information, 1991.

Keen, E.M. A retrieval comparison of six published indexes in the field of library and information science. *Unesco Bulletin for Libraries*, 30, 1976, 26-36.

Keen, E.M.; Digger, J.A. *Report of an information science index language test*. Aberystwyth, College of Librarianship Wales, 1972. 2 volumes.

Kehl, W.B. et al. An information retrieval language for legal studies. *Communications of the ACM*, 4, 1961, 380-389.

Keister, L.H. User types and queries: impact on image access systems. In: *Challenges in indexing electronic text and images*; ed. by R. Fidel et al., p. 7-22. Medford, NJ, Learned Information, 1994.

Kellman, S.G., ed. *Masterplots II: American fiction series*. 6 volumes. Pasadena, CA, Salem Press, 2000.

Kent, A. et al. Relevance predictability in information retrieval systems. *Methods of Information in Medicine*, 6, 1967, 45-51.

Kerpedjiev, S.M. Automatic generation of multimodal weather reports from datasets. In:*Proceedings of the Third Conference on Applied Natural Language Processing*, p. 48-55. San Francisco, Morgan Kaufmann, 1992.

Kessler, M.M. Bibliographic coupling between scientific papers. *American Documentation*, 14, 1963, 10-25.

Kessler, M.M. *Bibliographic coupling extended in time*. Cambridge, MA, Massachusetts, Institute of Technology, 1962.

Kessler, M.M. Comparison of results of bibliographic coupling and analytic subject indexing. *American Documentation*, 16, 1965, 223-233.

Kim, W.; Wilbur, W.J. Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology*, 52, 2001, 247-259.

King, R. A comparison of the readability of abstracts with their source documents. *Journal of the American Society for Information Science*, 27, 1976, 118-121.

Klement, S. Open-system versus closed-system indexing. *The Indexer*, 23, 2002, 23-31.

Klingbiel, P.H. *The future of indexing and retrieval vocabularies*. Alexandria, VA, Defense Documentation Center, 1970. AD 716 200.

Klingbiel, P.H. *Machine-aided indexing*. Technical progress report for period July 1969-June 1970. Alexandria, VA, Defense Documentation Center, 1971. AD 721 875.

Klingbiel, P.H. & Rinker, C.C. Evaluation of machine-aided indexing. *Information Processing and Management*, 12, 1976, 351-366.

Knapp, S.D. BRS/TERM: database for searchers. *Online '83 Conference Proceedings*, p. 162-166. Weston, CT, Online Inc., 1983.

Knapp, S.D. *The contemporary thesaurus of social science terms and synonyms: a guide for natural language computer searching*. Phoenix, AZ, Oryx Press, 1993.

Knapp, S.D. Free-text searching of online databases. *Reference Librarian*, 5/6, 1982, 143-153.

Knight, K. Mining online text. *Communications of the ACM*, 42(11), 1999, 58-61.

Knorz, G. *Automatisches Indexieren als Erkennen abstrakter Objekte*. Tübingen, Max Niemeyer Verlag, 1983.

Kolcz, A. Summarization as feature selection for text categorization. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, p. 365-370. New York, Association for Computing Machinery, 2001.

Korotkin, A.L.; Oliver, L.H. *The effect of subject matter familiarity and the use of an indexing aid upon inter-indexer consistency*. Bethesda, MD, General Electric Company, Information Systems Operation, 1964.

Korycinski, C.; Newell, A.F. Natural-language processing and automatic indexing. *The Indexer*, 17, 1990, 21-29.

Krause, M.G. Intellectual problems of indexing picture collections. *Audiovisal Librarian*, 14, 1988, 73-81.

Krieger, T. *Instructor influences versus text influences in the selection of subject descriptors by undergraduate students*. Doctoral thesis. Urbana-Champaign, University of Illinois, Graduate School of Library Science, 1981.

Kubala, F. et al. Integrated technologies for indexing spoken language. *Communications of the ACM*, 43(2), 2000, 48-56.

Kuhlen, R. Some similarities and differences between intellectual and machine text understanding for the purpose of abstracting. In: *Representation and exchange of knowledge as a basis of information processes*; ed. by H.J. Dietschmann, p. 87-109. Amsterdam, North-Holland, 1984.

Kupiec, J.M. Murax: finding and organizing answers from text search. In: *Natural language information retrieval*; ed. by T.S. Strzalkowski, p. 311-332. Boston, Kluwer, 1999.

Kurita, T.; Kato, T. Learning of personal visual impression for image database systems. In: *Proceedings of the Second International Conference on Document Analysis and Recognition*, p. 547-552. Los Alamitos, CA, IEEE Computer Society Press, 1993.

Kwok, K.L. A probabilistic theory of indexing and similarity measure based on cited and citing documents. *Journal of the American Society for Information Science*, 36, 1985a, 342-351.

Kwok, K.L. A probabilistic theory of indexing using author-provided relevance information. *Proceedings of the American Society for Information Science*, 22, 1985b, 59-63.

Kwon, O-W.; Lee, J.-H. Text categorization based on k-nearest neighbor approach for Web site classification. *Information Processing & Management*, 39, 2003, 25-44.

LaBorie, T. et al. Library and information science abstracting and indexing services: coverage, overlap, and context. *Library and Information Science Research*, 7, 1985, 183-195.

Lam-Adesina, A.M.; Jones, G.J.F. Applying summarization techniques for term selection in relevance feedback. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1-9. New York, Association for Computing Machinery, 2001.

Lancaster, F.W. *Evaluation of the MEDLARS demand search service*. Bethesda, MD, National Library of Medicine, 1968.

Lancaster, F.W. *Information retrieval systems: characteristics, testing and evaluation*. New York, Wiley, 1968b.

Lancaster, F.W. Some observations on the performance of EJC role indicators in a mechanized retrieval system. *Special Libraries*, 55, 1964, 696-701.

Lancaster, F.W. *Vocabulary control for information retrieval*. 2nd ed. Arlington, VA, Information Resources Press, 1986.

Lancaster, F.W. *Vocabulary control for information retrieval*. Washington, DC, Information Resources Press, 1972.

Lancaster, F.W.; Sandore, B. *Technology and management in library and information services*. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Lancaster, F.W.; Warner, A.J. *Information retrieval today*. Arlington, VA, Information Resources Press, 1983.

Lancaster, F.W.; Warner, A.J. *Intelligent technologies in library and information service applications*. Medford, NJ, Information Today, 2001.

Lancaster, F.W. et al. Evaluating the effectiveness of an on-line, natural language retrieval system. *Information Storage and Retrieval*, 8, 1972, 223-245.

Lancaster, F.W. et al. Evaluation of interactive knowledge-based systems: overview and design for empirical testing. *Journal of the American Society for Information Science*, 47, 1996, 57-69.

Lancaster, F.W. et al. Identifying barriers to effective subject access in library catalogs. *Library Resources & Technical Services*, 35, 1991, 377-391.

Lancaster, F.W. et al. *Modular content analyses*. Final report to the National Science Foundation. Washington, DC, Herner and Company, 1965.

Larson, R.R. Experiments in automatic Library of Congress classification. *Journal of the American Society for Information Science*, 43, 1992, 130-148.

Lawrence, S.; Giles, C.L. Accessibility of information on the Web. *Nature*, 400, 1999, 107-109.

Lawrence, S. et al. Digital libraries and autonomous citation indexing. *Computer*, 32(6), 1999, 67-71.

Lawson, M. et al. Automatic extracion of citations from the text of English-language patents — an example of template mining. *Journal of Information Science*, 22, 1996, 423-436.

Layne, S.S. Some issues in the indexing of images. *Journal of the American Society for Information Science*, 45, 1994, 583-588.

Layne, S.S. Subject access to art images. In: *Introduction to art image access*; ed. by M. Baca, p. 1-19. Los Angeles, Getty Research Institute, 2002.

Leacock, C. et al. Corpus-based statistical sense resolution. In: *Human language technology: proceedings of a workshop held at Plainsboro, New Jersey, March 21-24, 1993*, p. 260-263. San Francisco, Morgan Kaufmann, 1993.

Lehmam, A. Text structuration leading to an automatic summary system: RAFI. *Information Processing & Management*, 35, 1999, 181-191.

Leighton, H.V.; Srivastava, J. First 20 precision among World Wide Web Search Services (Search Engines). *Journal of the American Society for Information Science*, 50, 1999, 870-881.

Leininger, K. Interindexer consistency in PSYCINFO. *Journal of Librarianship and Information Science*, 32(1), 2000, 4-8.

Leonard, L.E. *Inter-indexer consistency and retrieval effectiveness: measurement of relationships*. Doctoral thesis. Urbana-Champaign, University of Illinois, Graduate School of Library Science, 1975.

Levinson, S.E. Speech recognition technology: a critique. *Proceedings of the National Academy of Sciences*, 92, 1995, 9953-9955.

Li, Y. et al. Semantic image retrieval through human subject segmentation and characterization. In: *Storage and retrieval for image and video databases V*; ed. by I.K. Sethi and R.C. Jain, p. 340-351. Bellingham, WA, International Society for Optical Engineering, 1997.

Liddy, E.D. How a search engine works. In: *Web of deception: misinformation on the Internet*; ed. by A.P. Mintz, p. 197-208. Medford, NJ, Information Today, 2002.

Liddy, E.D.; Jörgensen, C. Modeling information seeking behaviors in index use. *Proceedings of the American Society for Information Science*, 30, 1993a, 185-190.

Liddy, E.D.; Jörgensen, C. Reality check! Book index characteristics that facilitate information access. In: *Indexing, providing access to information*; ed. by N.C. Mulvany, p. 125-138. Port Aransas, TX, American Society of Indexers, 1993b.

Liddy, E.D. et al. Index quality study, part II: publishers' survey and qualitative assessment. In: *Indexing tradition and innovation*, p. 53-79. American Society of Indexers, 1990.

Lieberman, H. et al. Aria: an agent for annotating and retrieving images. *Computer*, 34(7), 2001, 57-62.

Lienhart, R. et al. Video abstracting. *Communications of the ACM*, 40(12), 1997, 55-62.

Lippincott, A. Issues in content-based music information retrieval. *Journal of Information Science*, 28, 2002, 137-142.

Liu, C.-C.; Tsai, P.-J. Content-based retrieval of MP3 music objects. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, p. 506-511. New York, Association for Computing Machinery, 2001.

Liu, W. et al. A media agent for automatically building a personalized semantic index of Web media objects. *Journal of the American Society for Information Science and Technology*, 52, 2001, 853-855.

Liu, Y.; Li, F. Semantic extraction and semantics-based annotation and retrieval for video databases. *Multimedia Tools and Applications*, 17, 2002, 5-20.

Loukopoulos, L. Indexing problems and some of their solutions. *American Documentation*, 17, 1966, 17-25.

Lu, C. et al. TheSys — a comprehensive thesaurus system for intelligent document analysis and text retrieval. In: *Proceedings of Third International Conference on Document Analysis and Recognition*. Volume 2, p. 1169-1173. Los Alamitos, CA, IEEE Computer Society Press, 1995.

Lu, G. Indexing and retrieval of audio: a survey. *Multimedia tools and applications*, 15, 2001, 269-290.

Luhn, H.P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2, 1958, 159-165.

Luhn, H.P. *Keyword-in-context index for technical literature (KWIC index)*. Yorktown Heights, NY, IBM Advanced Systems Development Division, 1959.

Luhn, H.P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1, 1957, 309-317.

Lunin, L. The development of a machine-searchable index-abstract and its application to biomedical literature. In: *Three Drexel information science research studies*; ed. by B. Flood, p. 47-134. Philadelphia, Drexel Press, 1967.

Lynch, C.A. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology*, 52, 2001, 12-17.

Lynch, M.F.; Petrie, J.H. A program suite for the production of articulated subject indexes. *Computer Journal*, 16, 1973, 46-51.

Ma, W.-Y.; Manjunath, B.S. A texture thesaurus for browsing large aerial photographs. *Journal of the American Society for Information Science*, 49, 1998, 633-648.

Maarek, Y.S. Automatically constructing simple help systems from natural language documentation. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 243-256. Hillsdale, NJ, Lawrence Erlbaum, 1992.

MacDougall, S. Signposts on the information superhighway: indexes and access. *Journal of Internet Cataloging*, 2(3/4), 2000, 61-79.

MacEwan, A. Where do you keep the dystopias? *Library Association Record*, 99, 1997, 40-41.

Magill, F.N., ed. *Masterplots: 2,010 plot stories & essay reviews from the world's fine literature*. Revised edition. Englewood Cliffs, NJ, Salem Press, 1976.

Magill, F.N., ed. *Masterplots II: American fiction series*. Volume 1. Englewood Cliffs, NJ, Salem Press, 1986.

Mai, J.-E. Deconstructing the indexing process. *Advances in Librarianship*, 23, 2000, 269-298.

Mai, J.-E. Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation*, 57, 2001, 591-622.

Malone, L.C. et al. Modeling the performance of an automated keywording system. *Information Processing & Management*, 27, 1991, 145-151.

Mani, I. *Automatic summarization*. Philadelphia, John Benjamins Publishing, 2001a.

Mani, I. Recent developments in text summarization. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, p. 529-531. New York, Association for Computing Machinery, 2001b.

Mani, I. et al. *TIPSTER SUMMAC text summarization evaluation. Final report.* MTR 98W0000138. McLean, VA, MITRE Corporation, 1998.

Mani, I. et al. Towards content-based browsing of broadcast news video. In: *Intelligent multimedia information retrieval*; ed. by M.T. Maybury, p. 241-258. Menlo Park, CA, AAAI Press, 1997.

Marchionini, G. et al. Extending retrieval strategies to networked environments: old ways, new ways, and a critical look at WAIS. *Journal of the American Society for Information Science*, 45, 1994, 561-564.

Marcus, R.S. et al. The user interface for the Intrex retrieval system. In: *Interactive bibliographic search: the user/computer interface*; ed. by D.E. Walker, p. 159-201. Montvale, NJ, AFIPS Press, 1971.

Markey, K. et al. An analysis of controlled vocabulary and free-text search statements in online searches. *Online Review*, 4, 1980, 225-236.

Markey, K. Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research*, 6, 1984, 155-177.

Markkula, M.; Sormunen, E. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1, 2000, 259-285.

Maron, M.E. Depth of indexing. *Journal of the American Society for Information Science*, 30, 1979, 224-228.

Maron, M.E. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, 28, 1977, 38-43.

Maron, M.E. Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing and Management*, 24, 1988, 249-250.

Maron, M.E.; Kuhns, J.C. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7, 1960, 216-244.

Maron, M.E. et al. *Probabilistic indexing — a statistical technique for document identification and retrieval*. Los Angeles, Thompson Ramo Wooldridge, 1959.

Marques, O.; Furht, B. MUSE: a content-based image search and retrieval system using relevance feedback. *Multimedia Tools and Applications*, 17, 2002, 21-50.

Marshall, C.C. The future of annotation in a digital (paper) world. In: *Successes & failures of digital libraries*; ed. by S. Harum and M. Twidale, p. 97-117. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 2000.

Martin, W.A. Toward an integral multi-file on-line bibliographic database. *Journal of Information Science*, 2, 1980, 241-253.

Martinez, C. et al. An expert system for machine-aided indexing. *Journal of Chemical Information and Computer Science*, 27, 1987, 158-162.

Martyn, J. Tests on abstracts journals: coverage, overlap, and indexing. *Journal of Documentation*, 23, 1967, 45-70.

Martyn, J.; Slater, M. Tests on abstracts journals. *Journal of Documentation*, 20, 1964, 212-235.

Massey-Burzio, V. The MultiPlatter experience at Brandeis University. *CD-ROM Professional*, 3(3), 1990, 22-26.

Mathis, B.A. *Techniques for the evaluation and improvement of computer-produced abstracts*. Columbus, Ohio State University, Computer and Information Science Research Center, 1972. OSU-CISRC-TR-72-15. PB 214 675.

Mathis, B.A. et al. Improvement of automatic abstracts by the use of structural analysis. *Journal of the American Society for Information Science*, 24, 1973, 101-109.

Maybury, M.T. Generating summaries from event data. *Information Processing & Management*, 31, 1995, 735-751.

McCain, K.W. et al. Comparing retrieval performance in online data bases. *Information Processing & Management*, 23, 1987, 539-553.

McCray, A.T. et al. Evaluating UMLS strings for natural language processing. *Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association*, p. 448-452. Philadelphia, Hanley & Belfus, 2001.

McDermott, J. Another analysis of full-text legal document retrieval. *Law Library Journal*, 78, 1986, 339-343.

McDonald, D.D. Robust partial-parsing through incremental, multi-algorithm processing. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 83-99. Hillsdale, NJ, Lawrence Erlbaum, 1992.

McDonald, S. et al. Evaluating a content based image retrieval system. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 232-240. New York, Association for Computing Machinery, 2001.

McKeown, K. et al. Generating concise natural language summaries. *Information Processing & Management*, 31, 1995, 703-733.

McNab, R.J. et al. Tune retrieval in the multimedia library. *Multimedia Tools and Applications*, 10, 2000, 113-132.

Medeiros, N. et al. Utilizing CORC to develop and maintain access to biomedical Web sites. In: *CORC: new tools and possibilities for cooperative electronic resource description*; ed. by K. Calhoun and J.J. Riemer, p. 111-121. Binghampton, NY, Haworth Press, 2001.

Mehrotra, R. Content-based image modeling and retrieval. In: *Digital image access & retrieval*; ed. by P.B. Heidorn and B. Sandore, p. 57-67. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Mehrotra, R.; Gary, J.E. Similar-shape retrieval in shape data management. *Computer*, 28(9), 1995, 57-62.

Mehtre, B.M. et al. Content-based image retrieval using a composite color-shape approach. *Information Processing & Management*, 34, 1998, 109-120.

Mehtre, B.M. et al. Shape measures for content based image retrieval: a comparison. *Information Processing & Management*, 33, 1997, 319-337.

Melucci, M. An evaluation of automatically constructed hypertexts for information retrieval. *Information Retrieval*, 1, 1999, 91-114.

Meng, W. et al. Concept hierarchy-based text database categorization. *Knowledge and Information Systems*, 4, 2002, 132-150.

Methods for examining documents, determining their subjects, and selecting indexing terms. Geneva, International Organization for Standardization, 1985. ISO 5963-1985 (E).

Milstead, J.L. *Subject access systems: alternatives in design*. Orlando, Academic Press, 1984.

Milstead, J.L.; Feldman, S. Metadata: cataloging by any other name... *Online*, 26(1), 1999, 66-74.

Mintz, A.P., ed. *Web of deception: misinformation on the Internet*. Medford, NJ, Information Today, 2002.

Missingham, R. Indexing the Internet: pinning jelly to the wall? LASIE, 27(3), 1996, 32-42.

Mitchell, S.; Mooney, M. INFOMINE. In: *The amazing Internet challenge*; ed. by A.T. Wells et al., p. 97-120. Chicago, American Library Association, 1999.

Mizzaro, S. Relevance: the whole history. In: *Historical studies in information science*; ed. by T.B. Hahn and M. Buckland, p. 221-244. Medford, NJ, Information Today, 1998.

Moens, M.-F. *Automatic indexing and abstracting of document texts*. Boston, Kluwer, 2000.

Moens, M.-F.; Dumortier, J. Text categorization: the assignment of subject descriptors to magazine articles. *Information Processing & Management*, 36, 2000a, 841-861.

Moens, M.-F.; Dumortier, J. Use of a text grammar for generating highlight abstracts of magazine articles. *Journal of Documentation*, 56, 2000b, 520-539.

Moens, M.-F. et al. Information extraction from legal texts: the potential of discourse analysis. *International Journal of Human-Computer Studies*, 51, 1999, 1155-1171.

Moghaddam, B. et al. Regions-of-interest and spatial layout for content-based image retrieval. *Multimedia Tools and Applications*, 14, 2001, 201-210.

Montague, B.A. Testing, comparison and evaluation of recall, relevance and cost of coordinate indexing with links and roles. *American Documentation* 16, 1965, 201-208.

Montgomery, R.R. An indexing coverage study of toxicological literature. *Journal of Chemical Documentation*, 13, 1973, 41-44.

Moreno, P.J. et al. From multimedia retrieval to knowledge management. *Computer*, 35(4), 2002, 58-59, 62-66.

Mostafa, J. Digital image representation and access. *Annual Review of Information Science and Technology*, 29, 1994, 91-135.

Mostafa, J.; Dillon, A. Design and evaluation of a user interface supporting multiple image query models. *Proceedings of the American Society for Information Science*, 33, 1996, 52-57.

Mowshowitz, A.; Kawaguchi, A. Bias on the Web. *Communications of the ACM*, 45(9), 2002, 56-60.

Muddamalle, M.R. Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. *Journal of the American Society for Information Science*, 49, 1998, 881-887.

Mullison, W.R. et al. Comparing indexing efficiency, effectiveness, and consistency with or without the use of roles. *Proceedings of the American Society for Information Science*, 6, 1969, 301-311.

Mulvany, N.C. *Indexing books*. Chicago, University of Chicago Press, 1994.

Munakata, T., ed. Knowledge discovery. *Communications of the ACM*, 42(11), 1999, 26-67.

Myers, J.M. Computers and the searching of law texts in England and North America: a review of the state of the art. *Journal of Documentation*, 29, 1973, 212-228.

Nakamura, Y. et al. Diagram understanding utilizing natural language text. In: *Proceedings of the Second International Conference on Document Analysis and Recognition*, p. 614-618. Los Alamitos, CA, IEEE Computer Society Press, 1993.

Nam, J.; Tewfik, A.H. Event-driven video extraction and visualization. *Multimedia Tools and Applications*, 16, 2002, 55-77.

Nasukawa, T.; Nagano, T. Text analysis and knowledge mining system. *IBM Systems Journal*, 40, 2001, 967-984.

Nielsen, H.J. The nature of fiction and its significance for classification and indexing. *Information Services & Use*, 17, 1997, 171-181.

Nomoto, T.; Matsumoto, Y. A new approach to unsupervised text summarization. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 26-34. New York, Association for Computing Machinery, 2001.

Oakman, R.L. The evolution of intelligent writing assistants: trends and future prospects. In: *Proceedings (of the) Sixth International Conference on Tools with Artificial Intelligence*, p. 233-234. Los Alamitos, CA, IEEE Computer Society Press, 1994.

O'Connor, B.C. *Explorations in indexing and abstracting: pointing, virtue, and power*. Englewood, CO, Libraries Unlimited, 1996.

O'Connor, B.C. et al. User reactions as access mechanism: an exploration based on captions for images. *Journal of the American Society for Information Science*, 50, 1999, 681-697.

O'Connor, J. Automatic subject recognition in scientific papers: an empirical study. *Journal of the Association for Computing Machinery*, 12, 1965, 490-515.

O'Connor, J.G.; Meadows, A.J. *Physics Abstracts* as a source of abstracts in astronomy. *Journal of Documentation*, 2, 1968, 107-112.

Odlyzko, A.M. Abstracting and reviewing in the digital era. *NFAIS Newsletter*, 41(6), 1999, 85, 90-92.

Odlyzko, A.M. Tragic loss or good riddance? The impending demise of traditional scholarly journals. *International Journal of Human-Computer Studies*, 42, 1995, 71-122.

Ogle, V.E.; Stonebraker, M. Chabot: retrieval from a relational database of images. *Computer*, 28(9), 1995, 40-48.

Oh, S.G. Document representation and retrieval using empirical facts. *Journal of the American Society for Information Science*, 49, 1998, 920-931.

Ojala, M. Web content extraction. *EContent*, 25(4), 2002, 39-41.

Olafsen, T.; Vokac, L. Authors' reply to R. Moss. *Journal of the American Society for Information Science*, 34, 1983, 294.

Olason, S.C. Let's get usable: usability studies for indexes. *The Indexer*, 22, 2000, 91-95.

Olderr, S. *Olderr's fiction subject headings: a supplement and guide to the LC Thesaurus*. Chicago, American Library Association, 1991.

Oliver, D.E.; Altman, R.B. Extraction of SNOMED concepts from medical record texts. In: *Eighteenth Annual Symposium on Computer Applications in Medical Care*, p. 179-183. Philadelphia, Hanley & Belfus, 1994.

Oliver, L.H. et al. *An investigation of the basic processes involved in the manual indexing of scientific documents*. Bethesda, MD, General Electric Co., Information Systems Operation, 1966. PB 169 415.

Olson, H.A.; Boll, J.J. *Subject analysis in online catalogs*. 2nd ed. Englewood, CO, Libraries Unlimited, 2001.

O'Neill, E.T.; Aluri, R. Library of Congress subject heading patterns in OCLC monographic records. *Library Resources & Technical Services*, 25, 1981, 63-80.

O'Neill, E.T. et al. Web characterization project: an analysis of metadata usage on the Web. *Journal of Library Administration*, 34, 2001, 359-374. Este artigo foi originalmente publicado na *Annual Review of OCLC Research 1998*, e o texto integral encontra-se disponível em linha em <http://www.oclc.org/research/publications/arr/>

Onyshkevych, B. Issues and methodology for template design for information extraction. In: *Proceedings of the Human Language Technology Workshop, March 8-11, 1994*, p. 171-176. San Francisco, Morgan Kaufmann, 1994.

Oppenheim, C. The patents coverage of *Chemical Abstracts*. *Information Scientist*, 8, 1974, 133-138.

Oppenheim, C. et al. The evaluation of WWW search engines. *Journal of Documentation*, 56, 2000, 190-211.

Orbach, B. So that others may see: tools for cataloging still images. *Cataloging & Classification Quarterly*, 11(3/4), 1990, 163-191.

Ornager, S. The image database: a need for innovative indexing and retrieval. *Advances in Knowledge Organization*, 4, 1994, 208-216.

Ornager, S. Image retrieval: theoretical analysis and empirical user studies on accessing information in images. *Proceedings of the American Society for Information Science*, 34, 1997, 202-211.

Oswald, V.A., Jr. et al. *Automatic indexing and abstracting of the contents of documents*. Los Angeles, Planning Research Corporation, 1959. RADC-TR-59-208.

Over, P. The TREC interactive track: an annotated bibliography. *Information Processing & Management*, 37, 2001, 369-381.

Owen, P. Structured for success: the continuing role of quality indexing in intelligent information retrieval systems. In: *Online Information 94*, p. 227-231. Medford, NJ, Learned Information, 1994.

Ozaki, K. et al. Semantic retrieval on art museum database system. In: (Proceedings of the) *1996 IEEE International Conference on Systems, Man and Cybernetics*, p. 2108-2112. Piscataway, NJ, Institute of Electrical and Electronics Engineers, 1996.

Paice, C.D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In: *Information retrieval research*; ed. by R.N. Oddy et al., p. 172-191. London, Butterworths, 1981.

Paice, C.D.; Jones, P.A. The identification of important concepts in highly structured technical papers. In: SIGIR-93: *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 69-78. New York, Association for Computing Machinery, 1993.

Pao, M.L. Term and citation searching: a preliminary report. *Proceedings of the American Society for Information Science*, 25, 1988, 177-180.

Pao, M.L.; Worthen, D.B. Retrieval effectiveness by semantic and citation searching. *Journal of the American Society for Information Science*, 40, 1989, 226-235.

Patel, N.V.; Sethi, I.K. Audio characterization for video indexing. In: *Storage and retrieval for still image and video databases IV*; ed. by I.K. Sethi and R.C. Jain, p. 373-384. Bellingham, WA, International Society for Optical Engineering, 1996.

Patel, N.V.; Sethi, I.K. Video classification using speaker identification. In: *Storage and retrieval for image and video databases V*; ed. by I.K. Sethi and R.C. Jain, p. 218-225. Bellingham, WA, International Society for Optical Engineering, 1997.

Patrick, T.B. et al. Text indexing of images based on graphical image content. *Proceedings of the American Society for Information Science*, 36, 1999, 675-680.

Payne, D. et al. *A textual abstracting technique: a preliminary development and evaluation support*. Pittsburgh, PA, American Institutes for Research, 1962. 2 volumes. AD 285081-285082.

Pazienza, M.T., ed. *Information extraction*. New York, Springer-Verlag, 1999.

Pejtersen, A.M. Design of a computer-aided user-system dialogue based on an analysis of users' search behaviour. *Social Science Information Studies*, 4, 1984, 167-183.

Pejtersen, A.M. A framework for indexing and representation of information based on work domain analysis: a fiction classification example. *Advances in Knowledge Organization*, 4, 1994, 251-263.

Pejtersen, A.M. The meaning of 'about' in fiction indexing and retrieval. *Aslib Proceedings*, 31, 1979, 251-257.

Pejtersen, A.M. New model for multimedia interfaces to online public access catalogues. *Electronic Library*, 10, 1992, 359-366.

Pejtersen, A.M.; Austin, J. Fiction retrieval: experimental design and evaluation of a search system based on users' value criteria. *Journal of Documentation*, 39, 1983, 230-246; 40, 1984, 25-35.

Pentland, A. Machine understanding of human behavior in video. In: Intelligent multimedia information retrieval; ed. by M.T. Maybury, p. 175-188. Menlo Park, CA, AAAI Press, 1997.

Pereira, F.C.N.; Grosz, B.J. *Natural language processing*. Cambridge, MA, MIT Press, 1994.

Perez, E. Text enhancement: controlled vocabulary vs. free text. *Special Libraries*, 73, 1982, 183-192.

Perrone, M.P. Machine learning in a multimedia document retrieval framework. *IBM Systems Journal*, 41, 2002, 494-503.

Perry, J.W.; Kent, A. *Tools for machine literature searching*. New York, Interscience Publishers Inc., 1958.

Petrarca, A.E.; Lay, W.M. The double-KWIC coordinate index: a new approach for preparation of high-quality printed indexes by automatic indexing techniques. *Journal of Chemical Documentation*, 9, 1969, 256-261.

Picard, R.W. A society of models for video and image libraries. *IBM Systems Journal*, 35, 1996, 296-312.

Picard, R.W.; Minka, T.P. Vision texture for annotation. *Multimedia Systems*, 3, 1995, 3-14.

Pickens, J. Feature selection for polyphonic music retrieval. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 428-429. New York, Association for Computing Machinery, 2001.

Pinto, M. Documentary abstracting: toward a methodological model. *Journal of the American Society for Information Science*, 46, 1995, 225-234.

Pinto, M. Interdisciplinary approaches to the concept and practice of written text documentary content analysis (WTDCA). *Journal of Documentation*, 50, 1994, 111-133.

Pinto, M. *El resumen documental*. 2.ed. Madrid, Fundación Germán Sánchez Ruipérez, 2001.

Pinto, M.; Gálvez, C. Paradigms for abstracting systems. *Journal of Information Science*, 25, 1999, 365-380.

Pinto, M.; Lancaster, F.W. Abstracts and abstracting in knowledge discovery. *Library Trends*, 48, 1999, 234-248.

Piternick, A. Searching vocabularies: a developing category of online search tools. *Online Review*, 8, 1984, 441-449.

Pitkin, R.M.; Branagan, M.A. Can the accuracy of abstracts be improved by providing specific instructions? *Journal of the American Medical Association (JAMA)*, 280, 1998, 267-269.

Pitkin, R.M. et al. Accuracy of data in abstracts of published research articles. *Journal of the American Medical Association (JAMA)*, 281, 1999, 1110-1111.

Pitkin, R.M. et al. Effectiveness of journal intervention to improve abstract quality. *Journal of the American Medical Association (JAMA)*, 283, 2000, 481.

Place, E. Social Science Information Gateway (SOSIG). In: *The amazing Internet challenge*; ed. by A.T. Wells et al., p. 223-244. Chicago, American Library Association, 1999.

Plaunt, C.; Norgard, B.A. An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science*, 49, 1998, 888-902.

Pozzi, C.; Celentano, A. Knowledge-based document filing. *IEEE Expert*, 8(5), 1993, 34-45.

Prabha, C. The large retrieval phenomenon. *Advances in Library Automation and Networking*, 4, 1991, 55-92.

Preschel, B.M. *Funk & Wagnalls new encyclopedia indexing manual*. New York, Funk & Wagnall, 1981. (Inédito.)

Preschel, B.M. *Indexer consistency in perception of concepts and in choice of terminology*. New York, Columbia University, School of Library Service, 1972.

Price, D.S. Possible impact of electronic publishing on abstracting and indexing. *Journal of the American Society for Information Science*, 34, 1983, 288.

Price, R. et al. Applying relevance feedback to a photo archival system. *Journal of Information Science*, 18, 1992, 203-215.

Proceedings of the Third International Conference on Document Analysis and Recognition. Los Alamitos, CA, IEEE Computer Society Press, 1995. 2 v.

Qin, J. Semantic similarities between a keyword database and a controlled vocabulary database. *Journal of the American Society for Information Science*, 51, 2000, 166-180.

Qin, J.; Norton, M., ed. Knowledge discovery in bibliographic databases. *Library Trends*, 48(1), 1999 (todo o fascículo).

Ragusa, J.M.; Turban, E. Integrating expert systems and multimedia: a review of the literature. *International Journal of Applied Expert Systems*, 2(1), 1994, 54-71.

Raitt, D. Recall and precision devices in interactive bibliographic search and retrieval systems. *Aslib Proceedings*, 32, 1980, 281-301.

Rajagopalan, R. The Figure Understander: a tool for the integration of text and graphical input to a knowledge base. In: *Proceedings (of the) Sixth International Conference on Tools with Artificial Intelligence*, p. 80-87. Los Alamitos, CA, IEEE Computer Society Press, 1994.

Ramsey, M.C. et al. A collection of visual thesauri for browsing large collections of geographic images. *Journal of the American Society for Information Science*, 50, 1999, 826-834.

Ranta, J.A. The new literary scholarship and a basis for increased subject catalog access to imaginative literature. *Cataloging & Classification Quarterly*, 14(1), 1991, 3-26.

Rapoza, J. A smart way to put help on the Web. *PC Week*, 13(39), 1996, 93.

Rasheed, M.A. Comparative index terms. *International Library Review*, 21, 1989, 289-300.

Rasmussen, E.M. Indexing images. *Annual Review of Information Science and Technology*, 32, 1997, 169-196.

Rath, G.J. et al. Comparison of four types of lexical indicators of content. *American Documentation*, 12, 1961a, 126-130.

Rath, G.J. et al. The formation of abstracts by the selection of sentences. *American Documentation*, 12, 1961b, 139-143.

Ravela, S.; Luo, C. Appearance-based global similarity retrieval of images. In: *Advances in information retrieval*; ed. by W.B. Croft, p. 267-303. Boston, Kluwer, 2000.

Reamy, T. Auto-categorization — coming to a library or intranet near you! *EContent*, 25(11), 2002, 16-22.

Reich, P.; Biever, E.J. Indexing consistency: the input/output function of thesauri. *College & Research Libraries*, 52, 1991, 336-342.

Reisner, P. *Evaluation of a 'growing' thesaurus*. Yorktown Heights, NY, IBM, Thomas Watson Research Center, 1966. Research paper RD-1662.

Resnick, A. Relative effectiveness of document titles and abstracts for determining relevance of documents. *Science*, 134, 1961, 1004-1006.

Resnikoff, H.L.; Dolby, J.L. *Access: a study of information storage and retrieval with emphasis on library information systems*. 1972. ERIC Document ED 060 921.

Ribeiro-Neto, B. et al. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology*, 52, 2001, 391-401.

Rickman, R.M.; Stonham, T.J. Image retrieval from large databases using a neural network coding scheme. In: *The structuring of information: informatics II*; ed. by K.P. Jones, p. 147-159. London, Aslib, 1991.

Riloff, E.; Lehnert, W. Automated dictionary construction for information extraction from text. In: *Proceedings (of the) Ninth Conference on Artificial Intelligence for Applications*, p. 93-99. Los Alamitos, CA, IEEE Computer Society Press, 1993.

Rindflesch, T.C.; Aronson, A.R. Ambiguity resolution while mapping free text to the UMLS metathesaurus. In: *Eighteenth Annual Symposium on Computer Applications in Medical Care*, p. 240-244. Philadelphia, PA, Hanley & Belfus, 1994.

Rindflesch, T.C. et al. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing*, 5, 2000a, 514-525.

Rindflesch, T.C. et al. Extracting molecular binding relationships from biomedical text. *Proceedings of the Sixth Conference on Applied Natural Language Processing*, p. 188-195. San Francisco, CA, Morgan Kaufmann, 2000b.

Rindflesch, T.C. et al. Mining molecular binding terminology from biomedical text. *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association*, p. 127-131. Philadelphia, Hanley & Belfus, 1999.

Ro, J.S. An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I. On the effectiveness of full-text retrieval. *Journal of the American Society for Information Science*, 39, 1988, 73-78.

Roberts, D.; Souter, C. The automation of controlled vocabulary subject indexing of medical journal articles. *Aslib Proceedings*, 52, 2000, 384-400.

Robertson, S.E. Introduction to the special issue: overview of the TREC routing and filtering tasks. *Information Retrieval*, 5, 2002, 127-137.

Robertson, S.E. The parametric description of retrieval tests. *Journal of Documentation*, 25, 1969, 1-27, 93-107.

Robinson, J.; Hu, M. DOE's Energy Database (EDB) versus other energy related databases: a comparative analysis. *Database*, 4(4), 1981, 10-27.

Rodgers, D.J. *A study of inter-indexer consistency*. Washington, DC, General Electric Co., 1961.

Rolling, L. Indexing consistency, quality and efficiency. *Information Processing & Management*, 17, 1981, 69-76.

Rowe, N.C. Inferring depictions in natural-language captions for efficient access to picture data. *Information Processing & Management*, 30, 1994, 379-388.

Rowe, N.C. *Precise and efficient access to captioned picture libraries: the MARIE project*. Monterrey, CA, Naval Postgraduate School, Computer Science Department, 1996.

Rowe, N.C.; Frew, B. Automatic caption localization for photographs on World Wide Web pages. *Information Processing & Management*, 34, 1998, 95-107.

Rowe, N.C.; Frew, B. Automatic classification of objects in captioned depictive photographs for retrieval. In: *Intelligent multimedia information retrieval*; ed. by M. Maybury, p. 65-79. Palo Alto, CA, AAAI Press, 1997.

Rowe, N.C.; Guglielmo, E.J. Exploiting captions in retrieval of multimedia data. *Information Processing & Management*, 29, 1993, 453-461.

Runde, C.E.; Lindberg, W.H. The curse of Thamus: a response. *Law Library Journal*, 78, 1986, 345-347.

Rush, J.E. et al. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22, 1971, 260-274.

Saarti, J. Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation*, 58, 2002, 49-65.

Saarti, J. Fiction indexing and the development of fiction thesauri. *Journal of Librarianship and Information Science*, 31, 1999, 85-92.

Saarti, J. Fiction indexing by library professionals and users. *Scandinavian Public Library Quarterly*, 33(4), 2000a, 6-9.

Saarti, J. Taxonomy of novel abstracts based on empirical findings. *Knowledge Organization*, 27, 2000b, 213-220.

Saggion, H.; Lapalme, G. Selective analysis for the automatic generation of summaries. In: *Dynamism and stability in knowledge organization*; ed. by C. Beghtol et al., p. 176-181. Würzburg, ERGON Verlag, 2000.

Salager-Meyer, F. Medical English abstracts: how well are they structured? *Journal of the American Society for Information Science*, 42, 1991, 528-531.

Salisbury, B.A., Jr.; Stiles, H.E. The use of the B-coefficient in information retrieval. *Proceedings of the American Society for Information Science*, 6, 1969, 265-268.

Salton, G. Another look at automatic text-retrieval systems. *Communications of the ACM*, 29, 1986, 648-656.

Salton, G. *Dynamic information and library processing*. Englewood Cliffs, NJ, Prentice-Hall, 1975.

Salton, G. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*, 23, 1972, 75-84.

Salton, G. *A syntactic approach to automatic book indexing*. Ithaca, NY, Cornell University, Department of Computer Science, 1989. Technical Report TR 89-979.

Salton, G.; Buckley, C. Automatic text structuring experiments. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 199-210. Hillsdale, NJ, Lawrence Erlbaum, 1992.

Salton, G.; McGill, M.J. *Introduction to modern information retrieval*. New York, McGraw Hill, 1983.

Salton, G.; Zhang, Y. Enhancement of text representations using related document titles. *Information Processing & Management*, 22, 1986, 385-394.

Salton, G. et al. Automatic text structuring and summarization. *Information Processing & Management*, 33, 1997, 193-207.

Santini, S. Using language more responsibly. *Computer*, 35(12), 2002, 126-128.

Sapp, G. The levels of access: subject approaches to fiction. *RQ*, 25, 1986, 488-497.

Saracevic, T. Comparative effects of titles, abstracts and full texts on relevance judgements. *Proceedings of the American Society for Information Science*, 6, 1969, 293-299.

Saracevic, T. et al. Letter to the editor. *Information Processing & Management*, 39, 2003, 153-156.

Saracevic, T. et al. A study of information seeking and retrieving. *Journal of the American Society for Information Science*, 39, 1988, 161-216.

Šauperl, A. *Subject determination during the cataloging process*. Lanham, MD, Scarecrow Press, 2002.

Savić, D. Automatic classification of office documents: review of available methods and techniques. *Records Management Quarterly*, 29(4), 1995, 3-6, 8-18.

Savoy, J. A new probabilistic scheme for information retrieval in hypertext. *New Review of Hypermedia and Multimedia*, 1, 1995, 107-134.

Schiffman, B. et al. Producing biographical summaries. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 450-457. New Brunswick, NJ, Association for Computational Linguistics, 2001.

Schreiber, A.Th. et al. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3), 2001, 66-74.

Schroeder, K.A. Layered indexing of images. *The Indexer*, 21, 1998, 11-14.

Schwarz, C. Web search engines. *Journal of the American Society for Information Science*, 49, 1998, 973-982.

Scott, D.W. Museum Data Bank research report. *Library Trends*, 37, 1988, 130-141.

Sekerak, R.J. A comparison of journal coverage in *Psychological Abstracts* and the primary health sciences indexes: implications for cooperative serials acquisition and retention. *Bulletin of the Medical Library Association*, 74, 1986, 231-233.

Seloff, G.A. Automated access to the NASA-JSC image archives. *Library Trends*, 38, 1990, 682-696.

Selye, H. *Symbolic shorthand system*. New Brunswick, NJ, Rutgers State University, Graduate School of Library Service, 1966.

Selye, H.; Ember, G. *Symbolic shorthand system for physiology and medicine*. 4th ed. Montreal, Université de Montreal, 1964.

Semeraro, G. et al. Learning contextual rules for document understanding. In: *Proceedings* (of the) *Tenth Conference on Artificial Intelligence for Applications*, p. 108-115. Los Alamitos, CA, IEEE Computer Society Press, 1994.

Shafer, K.E. Evaluating Scorpion results. *Journal of Library Administration*, 34, 2001, 237-244.

Shafer, K.E. Scorpion helps catalog the Web. *Bulletin of the American Society for Information Science*, 24(1), 1997, 28-29.

Sharp, J.R. The SLIC index. *American Documentation*, 17, 1966, 41-44.

Shatford, S. Analyzing the subject of a picture: a theoretical approach. *Cataloging & Classification Quarterly*, 6(3), 1986, 39-62.

Shaw, W.M., Jr. An investigation of document partitions. *Information Processing & Management*, 22, 1986, 19-28.

Shaw, W.M., Jr. An investigation of document structures. *Information Processing & Management*, 26, 1990a, 339-348.

Shaw, W.M., Jr. Subject indexing and citation indexing. *Information Processing & Management*, 26, 1990b, 693-718.

Shirey, D.L.; Kurfeerst, M. Relevance predictability: II. Data reduction. In: *Electronic handling of information: testing and evaluation*; ed. by A. Kent et al., p. 187-198. Washington, DC, Thompson Book Co., 1967.

Shneiderman, B. The limits of speech recognition. *Communications of the ACM*, 43(9), 2000, 63-65.

Shuldberg, H.K. et al. Distilling information from text: the EDS TemplateFiller system. *Journal of the American Society for Information Science*, 44, 1993, 493-507.

Sievert, M.; McKinin, E.J. Why full-text misses some relevant documents: an analysis of documents not retrieved by CCML or MEDIS. *Proceedings of the American Society for Information Science*, 26, 1989, 34-39.

Sievert, M. et al. Retrieval from full-text medical literature: the dream & the reality. In: *Fifteenth Annual Symposium on Computer Applications in Medical Care*, p. 348-352. New York, McGraw Hill, 1992.

Silvester, J.P. Computer supported indexing. In: *Encyclopedia of library and information science*. Volume 61, Supplement 24, p. 76-90. New York, Marcel Dekker, 1998.

Silvester, J.P. et al. Machine-aided indexing at NASA. *Information Processing & Management*, 30, 1994, 631-645.

Silvester, J.P. et al. Machine aided indexing from natural language text. Status report. Linthicum Heights, MD, RMS Associates, 1993. NASA-CR-4512.

Singhal, A.; Pereira, F. Document expansion for speech retrieval. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, p. 34-41. New York, Association for Computing Machinery, 1999.

Sinnett, J.D. *An evaluation of links and roles used in information retrieval*. Dayton, Air Force Materials Laboratory, Wright Patterson Air Force Base, 1964. AD 432 198.

Slamecka, V.; Jacoby, J. *Effect of indexing aids on the reliability of indexers. Final technical note*. Bethesda, MD, Documentation Inc., 1963. RADC-TDR-63-116.

Small, H. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 1973, 265-269.

Smalley, T.N. Comparing *Psychological Abstracts* and *Index Medicus* for coverage of the journal literature in a subject area in psychology. *Journal of the American Society for Information Science*, 31, 1980, 143-146.

Smeaton, A.F. Using NLP or NLP resources for information retrieval tasks. In: *Natural language information retrieval*; ed. by T. Strzalkowski, p. 99-111. Boston, Kluwer, 1999.

Smith, F.J. et al. Voice access to BLAISE. In: *Online Information 89*, p. 1-12. Medford, NJ, Learned Information, 1989.

Smith, G.L. Generation of eletronic product documentation. In: Innovative applications of artificial intelligence 2; ed. by A. Rappaport and R. Smith, p. 189-200. Cambridge, MA, MIT Press, 1991.

Smith, J.R.; Chang, S.-F. An image and video search engine for the World-Wide Web. In: *Storage and retrieval for image and video databases V*; ed. by I.K. Sethi and R.C. Jain, p. 84-95. Bellingham, WA, International Society for Optical Engineering, 1997a.

Smith, J.R.; Chang, S.-F. Querying by color regions using the VisualSEEK content-based visual query system. In: *Intelligent multimedia information retrieval*; ed. by M.T. Maybury, p. 23-41. Menlo Park, CA, AAAI Press 1997b.

Sneiderman, C.A. et al. Identification of anatomical terminology in medical text. *Proceedings of the 1998 Annual Symposium of the American Medical Informatics Association*, p. 428-432. Philadelphia, Hanley & Belfus, 1998.

Snow, B. et al. Grateful MED: NLM's front end software. *Database*, 9(6), 1986, 94-99.

Soergel, D. *Indexing languages and thesauri: construction and maintenance*. Los Angeles, CA, Melville, 1974.

Soergel, D. *Organizing information: principles of data base and retrieval systems*. Orlando, Academic Press, 1985.

Soergel, D. The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, 50, 1999, 1119-1120.

Solov'ev, V.I. The aspective method of abstracting. *Automatic Documentation and Mathematical Linguistics*, 5(1), 1971, 30-35. (Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 2, número 2, 1971, p. 14-17.)

Solov'ev, V.I. Functional characteristics of the author's abstract of a dissertation and the specifics of writing it. *Scientific and Technical Information Processing*, 3, 1981, 80-88. (Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 1, número 6, 1981, p. 20-24.)

Sparck Jones, K. Does indexing exhaustivity matter? *Journal of the American Society for Information Science*, 24, 1973, 313-316.

Sparck Jones, K. Letter to the editor. *Information Processing & Management*, 39, 2003, 156-159.

Sparck Jones, K. Reflections on TREC. *Information Processing & Management*, 31(3), 1995, 291-314.

Sparck Jones, K. What is the role of NLP in text retrieval? In: *Natural language information retrieval*; ed. by T. Strzalkowski, p. 1-24. Boston, Kluwer, 1999.

Spinellis, D. The decay and failures of Web references. *Computer*, 46(1), 2003, 71-77.

Srihari, R.K. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9), 1995a, 49-56.

Srihari, R.K. Automatic indexing and content-based retrieval of captioned photographs. In: *Proceedings of Third International Conference on Document Analysis and Recognition*. Volume 2, p. 1165-1167. Los Alamitos, CA, IEEE Computer Society Press, 1995b.

Srihari, R.K. Intelligent document understanding: understanding photographs with captions. In: *Proceedings of the Second International Conference on Document Analysis and Recognition*, p. 664-667. Los Alamitos, CA, IEEE Computer Society Press, 1993.

Srihari, R.K. Using speech input for image interpretation, annotation, and retrieval. In: *Digital image access & retrieval*; ed. by P.B. Heidorn and B. Sandore, p. 140-156. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Srinavasan, P. et al. An investigation of indexing on the WWW. *Proceedings of the American Society for Information Science*, 33, 1996, 79-83.

Srinavasan, S.; Brown, E.W. Is speech recognition becoming mainstream? *Computer*, 35(4), 2002, 38-41.

Srinavasan, S.; Petkovic, D. Phonetic confusion matrix-based spoken document retrieval. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 81-87. New York, Association for Computing Machinery, 2001.

Stanfill, C.; Waltz, D.L. Statistical methods, artificial intelligence, and information retrieval. In: *Text-based intelligent systems*; ed. P.S. Jacobs, p. 215-225. Hillsdale, NJ, Lawrence Erlbaum, 1992.

Stiles, H.E. Machine retrieval using the association factor. In: *Machine indexing: progress and problems*, p. 192-206. Washington, DC, American University, 1961.

Stock, O. ALFRESCO: enjoying the combination of natural language processing and hypermedia for information exploration. In: *Intelligent multimedia interfaces*; ed. by M.T. Maybury, p. 197-224. Cambridge, MA, MIT Press, 1993.

Stock, O. et al. Explorations in an environment for natural-language multimodal information access. In: *Intelligent Multimedia Information Retrieval*; ed. by M.T. Maybury, p. 381-398. Menlo Park, CA, AAAI Press, 1997.

Strzalkowski, T. et al. Evaluating natural language processing techniques in information retrieval. In: *Natural language information retrieval*; ed. by T. Strzalkowski, p. 113-145. Boston, Kluwer, 1999.

Stubbs, E.A. et al. Internal quality audit of indexing: a new application of interindexer consistency. *Cataloging and Classification Quarterly*, 28(4), 1999, 53-69.

Studwell, W.E. USE, the Universal Subject Environment: a new subject access approach in the time of the Internet. *Journal of Internet Cataloging*, 2(3/4), 1998, 197-209.

Su, L.T.; Chen, H.-I. Evaluation of Web search engines by undergraduate students. *Proceedings of the American Society for Information Science*, 36, 1999, 98-114.

Sundheim, B.M. Overview of results of the MUC-6 evaluation. In: *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, p. 13-31. San Francisco, Morgan Kaufmann, 1995.

Sutcliffe, A. et al. Empirical studies in multimedia information retrieval. In: *Intelligent multimedia information retrieval*; ed. by M.T. Maybury, p. 449-471. Menlo Park, CA, AAAI Press, 1997.

Sutton, S.A. Conceptual design and deployment of a metadata framework for educational resources on the Internet. *Journal of the American Society for Information Science*, 50, 1999, 1182-1192.

Svenonius, E. Access to nonbook materials: the limits of subject indexing for visual and aural languages. *Journal of the American Society for Information Science*, 45, 1994, 600-606.

Swanson, D.R. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association*, 78, 1990, 29-37.

Swanson, D.R. Searching natural language text by computer. *Science*, 132, 3434, 1960, 1099-1104.

Swanson, D.R. Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly*, 56, 1986, 389-398.

Swift, D.F. et al. 'Aboutness' as a strategy for retrieval in the social sciences. *Aslib Proceedings*, 30, 1978, 182-187.

Taddio, A. et al. Quality of nonstructured and structured abstracts of original research articles in the *British Medical Journal*, the *Canadian Medical Association Journal* and the *Journal of the American Medical Association. Canadian Medical Association Journal*, 150, 1994, 1611-1615.

Takeshita, A. et al. Topic-based multimedia structuring. In: *Intelligent multimedia information retrieval*; ed. by M.T. Maybury; p. 259-277. Menlo Park, CA, AAAI Press, 1997.

Tancredi, S.A.; Nichols, O.D. Air pollution technical information processing — the microthesaurus approach. *American Documentation*, 19, 1968, 66-70.

Tell, B.V. Document representation and indexer consistency. *Proceedings of the American Society for Information Science*, 6, 1969, 285-292.

Tenopir, C. *Retrieval performance in a full text journal article database*. Doctoral thesis. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1984. (Versões condensadas apareceram como: Tenopir, C. Full text database retrieval performance. *Online Review*, 9, 1985, 149-164 e Tenopir, C. Searching *Harvard Business Review. Online*, 9(2), 1985, 1-8.)

Tessier, J.A. Hypertext linking as a model of expert indexing. *Advances in Classification Research*, 2, 1992, 171-178.

Thé, L. Morph your help desk into customer support. *Datamation*, 42, January 15, 1996, 52-54.

Thelwall, M. A survey of search engine capabilities useful in data mining. *Proceedings of the American Society for Information Science and Technology*, 38, 2001, 24-29.

Thompson, C.W.N. The functions of abstracts in the initial screening of technical documents by the user. *Journal of the American Society for Information Science*, 24, 1973, 270-276.

Thompson, R. et al. Evaluating Dewey concepts as a knowledge base for automatic subject assignment. http://orc.rsch.oclc.org:6109/eval_c.html February 12, 1997.

Thorpe, P. An evaluation of *Index Medicus* in rheumatology: coverage, currency, and efficiency. *Methods of Information in Medicine*, 13, 1974, 44-47.

Tibbo, H.R. Abstracting across the disciplines: a content analysis of abstracts from the natural sciences, the social sciences, and the humanities with implications for abstracting standards and online information retrieval. *Library and Information Science Research*, 14, 1992, 31-56.

Tibbo, H.R. Indexing for the humanities. *Journal of the American Society for Information Science*, 45, 1994, 607-619.

Tinker, J.F. Imprecision in indexing. *American Documentation*, 17, 1966, 93-102; 19, 1968, 322-330.

Todeschini, C. Comunicação pessoal, 11 de novembro de 1997.

Todeschini, C.; Farrel, M.P. An expert system for quality control in bibliographic databases. *Journal of the American Society for Information Science*, 40, 1989, 1-11.

Todeschini, C.; Tolstenkov, A. Expert system for quality control in the INIS database. Paper presented at the International Symposium on the Future of Scientific, Technological and Industrial Information Services, Leningrad, May 1990. IAEA-SM-317/58.

Tong, R.M. et al. *RUBRIC: an environment for full text information retrieval*. Mountain View, CA, Advanced Information and Decision Systems, 1985.

Torr, D.V. et al. *Program of studies on the use of published indexes*. Bethesda, MD, General Electric Co., Information Systems Operation, 1966.

Trant, J. Framing the picture: standards for imaging systems. In: *Multimedia computing and museums*; ed. by D. Bearman, p. 347-367. Pittsburgh, Archives & Museum Informatics, 1995.

Trawinski, B. A methodology for writing problem structured abstracts. *Information Processing & Management*, 25, 1989, 693-702.

Trippe, B. Taxonomies and topic maps: categorization steps forward. *EContent*, 24(6), 2001, 44-49.

Troitskii, V.P. An extrapolation approach to the concept of information. *Automatic Documentation and Mathematical Linguistics*, 13(6), 1979, 49-60. (Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 2, 13(12), 1979, 1-7.)

Troitskii, V.P. Text, information and epistemology. *Automatic Documentation and Mathematical Linguistics*, 15(1), 1981, 20-27. (Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 2, 15(2), 1981, 1-5.)

Trubkin, L. Auto-indexing of the 1971-77 ABI/INFORM database. *Database*, 2(2), 1979, 56-61.

Trybula, W.J. Text mining. *Annual Review of Information Science and Technology*, 34, 1999, 385-419.

Tse, T. et al. An exploratory study of video browsing user interface designs and research methodologies. *Proceedings of the American Society for Information Science*, 36, 1999, 681-692.

Turner, J.M. Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: research results. *Proceedings of the American Society for Information Science*, 32, 1995, 9-12.

Turner, J.M. Representing and assessing information in the stockshot database at the National Film Board of Canada. Canadian Journal of Information Science, 15(4), 1990, 1-19.

Uhlmann, W. A thesaurus *Nuclear Science and Technology*: principles of design. *Teknisk-Vetenskaplig Forskning* (TVF), 38, 1967, 46-52.

Uthurusamy, R. et al. Extracting knowledge from diagnostic databases. *IEEE Expert*, 8(6), 1993, 27-38.

Vailaya, A. et al. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10, 2001, 117-130.

van der Meij, H. Styling the index: is it time for a change? *Journal of Information Science*, 28, 2002, 243-251.

Van der Meulen, W.A.; Janssen, P.J.F.C. Automatic versus manual indexing. *Information Processing & Management*, 13, 1977, 13-21.

van der Starre, J.H.E. Ceci n'est pas une pipe: indexing of images. In: *Multimedia computing and museums*; ed. by D. Bearman, p. 267-277. Pittsburgh, Archives & Museum Informatics, 1995.

Van Oot, J.G. et al. Links and roles in coordinate indexing and searching: an economic study of their use, and an evaluation of their effect on relevance and recall. *Journal of Chemical Documentation*, 6, 1966, 95-101.

Varney, S. Link your help desk to the Web. *Datamation*, 42(10), 1996, 64-67.

Vickery, B.C. The structure of semantic coding: a review. *American Documentation*, 10, 1959, 234-241.

Villarroel, M. et al. Obtaining feedback for indexing from highlighted text. *The Electronic Library*, 20, 2002, 306-313.

Vinsonhaler, J.F. Some behavioral indices of the validity of document abstracts. *Information Storage and Retrieval*, 3, 1966, 1-11.

Virgo, J.A. An evaluation of *Index Medicus* and MEDLARS in the field of ophthalmology. *Journal of the American Society for Information Science*, 21, 1970, 254-263.

Vizine-Goetz, D. Dewey in CORC: classification in metadata and pathfinders. In: *CORC: new tools and possibilities for cooperative electronic resource description*; ed. by K. Calhoun and J.J. Riemer, p. 67-80. Binghampton, NY, Haworth Press, 2001.

Vizine-Goetz, D. OCLC investigates using classification tools to organize Internet data. In: *Visualizing subject access for 21st century information resources*; ed. by P.A. Cochrane and E.H. Johnson, p. 93-105. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1998.

Vleduts-Stokolov, N. Concept recognition in an automatic text-processing system for the life sciences. *Journal of the American Society for Information Science*, 38, 1987, 269-287.

Voorbij, H.J. Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation*, 54, 1998, 466-476.

Voorhees, E.M. Natural language processing and information retrieval. In: *Information extraction*; ed. by M.T. Pazienza, p. 32-48. New York, Springer-Verlag, 1999.

Voorhees, E.M. Question answering in TREC. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, p. 535-537. New York, Association for Computing Machinery, 2001.

Voorhees, E.M.; Harman, D. The Text Retrieval Conferences (TRECs). In: *Proceedings of the TIPSTER Text Program, Phase III*, p. 241-267. San Francisco, Morgan Kaufmann, 1999.

Wactlar, H.D.; Christel, M.G. Digital video archives: managing through metadata. In: *Building a national strategy for digital preservation: issues in digital media archiving*, p. 80-95. Washington, DC, Council on Library and Information Resources, 2002.

Wactlar, H.D. et al. Complementary video and audio analysis for broadcast news archives. *Communications of the ACM*, 43(2), 2000, 42-47.

Wactlar, H.D. et al. Lessons learned from building a terabyte digital video library. *Computer*, 32(2), 1999, 66-73.

Walker, R.S. Problem child: some observations on fiction, with a sketch of a new system of classification. *Librarian and Book World*, 47(2), 1958, 21-28.

Walsh, J. Intel LANDesk lets users cry for help from Web browsers. *InfoWorld*, 18(39), 1996, 12.

Wang, J.Z. *Integrated region-based image retrieval*. Boston, MA, Kluwer, 2001.

Wanger, J. et al. *Evaluation of the on-line process*. Santa Monica, CA, Cuadra Associates, 1980. PB81-132565.

Watters, C. Information retrieval and the virtual document. *Journal of the American Society for Information Science*, 50, 1999, 1028-1029.

Watters, C.; Wang, H. Rating news documents for similarity. *Journal of the American Society for Information Science*, 51, 2000, 793-804.

Wechsler, M. et al. New approaches to spoken document retrieval. *Information Retrieval*, 3, 2000, 173-188.

Weeber, M. et al. Developing a test collection for biomedical word sense disambiguation. *Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association*, p. 746-750. Philadelphia, Hanley & Belfus, 2001.

Weil, B.H. et al. Technical-abstracting fundamentals. *Journal of Chemical Documentation*, 3, 1963, 86-89, 125-136.

Weinberg, B.H. Complexity in indexing systems — abandonment and failure: implications for organizing the Internet. *Proceedings of the American Society for Information Science*, 33, 1996, 84-90.

Weinberg, B.H. A theory of relativity for catalogers. In: *Cataloging heresy: challenging the standard bibliographic product*; ed. by B.H. Weinberg, p. 7-11. Medford, NJ, Learned Information, 1992.

Weinberg, B.H. Why indexing fails the researcher. *The Indexer*, 16, 1988, 3-6.

Weinberg, B.H. Why postcoordination fails the searcher. *The Indexer*, 19, 1995, 155-159.

Weld, D.S. et al., ed. The role of intelligent systems in the National Information Infrastructure. *AI Magazine*, 16(3), 1995, 45-64.

Wellisch, H.H. Book and periodical indexing. *Journal of the American Society for Information Science*, 45, 1994, 620-627.

Wells, A.T. et al. *The amazing Internet challenge*. Chicago, American Library Association, 1999.

Westberg, S. Comunicação transmitida por fax, em 9 de outubro de 1997.

Wheatley, A.; Armstrong, C.J. Metadata, recall, and abstracts: can abstracts ever be reliable indicators of document value? *Aslib Proceedings*, 49(8), 1997, 206-213.

White, H.D.; Griffith, B.C. Quality of indexing in online data bases. *Information Processing & Management*, 23, 1987, 211-224.

Wilbur, W.J. et al. Analysis of biomedical text for chemical names. *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association*, p. 176-180. Philadelphia, Hanley & Belfus, 1999.

Wilkinson, D.; Hollander, S. A comparison of drug literature coverage by *Index Medicus* and *Drug Literature Index*. *Bulletin of the Medical Library Association*, 61, 1973, 431-432.

Wilks, Y. et al. Combining weak methods in large-scale text processing. In: *Text-based intelligent systems*; ed. by P.S. Jacobs, p. 35-58. Hillsdale, NJ, Lawrence Erlbaum, 1992.

Williams, M. An evaluation of passage-level indexing strategies for a technical report archive. *LIBRES: Library and Information Science Electronic Journal*, v. 8, issue 1, March 31, 1998 (www.infomotions.com/serials/libres/libres-v8no1-williams-evaluation.txt)

Williams, M.E. Experiences of IIT Research Institute in operating a computerized retrieval system for searching a variety of data bases. *Information Storage and Retrieval*, 8, 1972, 57-75.

Wilson, P. Situational relevance. *Information Storage and Retrieval*, 9, 1973, 457-471.
Wilson, P. *Two kinds of power: an essay on bibliographical control*. Berkeley, University of California Press, 1968.
Winkler, M.A. The need for concrete improvement in abstract quality. *Journal of the American Medical Association (JAMA)*, 281, 1999, 1129-1130.
Witbrock, M.J.; Hauptmann, A.G. Speech recognition for a digital video library. *Journal of the American Society for Information Science*, 49, 1998, 619-632.
Wolfram, D. Inter-record linkage structure in a hypertext bibliographic retrieval system. *Journal of the American Society for Information Science*, 47, 1996, 765-774.
Wolfram, D.; Zhang, J. An investigation of the influence of indexing exhaustivity and term distributions on a document space. *Journal of the American Society for Information Science and Technology*, 53, 2002, 943-952.
Wong, K.-F. et al. Application of aboutness to functional benchmarking in information retrieval. *ACM Transactions on Information Systems*, 19, 2001, 337-370.
Wood, J.L. et al. Overlap among the journal articles selected for coverage by BIOSIS, CAS, and EI. *Journal of the American Society for Information Science*, 24, 1973, 25-28.
Wood, J.L. et al. Overlap in the lists of journals monitored by BIOSIS, CAS, and EI. *Journal of the American Society for Information Science*, 23, 1972, 36-38.
Woodland, P.C. et al. Effects of out of vocabulary words in spoken document retrieval. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 372-374. New York, Association for Computing Machinery, 2000
Woodruff, A.G.; Plaunt, C. GIPSY: automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45, 1994, 645-655.
Wooster, H. Optimal utilization of indexing personnel. *Research Review* (U.S. Air Force, Office of Aerospace Research), 3(4), 1964, 22-23.
Wright, L.W. et al. Hierarchical concept indexing of full-text documents in the Unified Medical Language System Information Sources Map. *Journal of the American Society for Information Science*, 50, 1999, 514-523.
Wu, J.K. et al. CORE: a content-based retrieval engine for multimedia information systems. *Multimedia Systems*, 3, 1995, 25-41.
Wu, J.K. et al. *Perspectives on content-based multimedia systems*. Boston, Kluwer, 2000.
Wu, Q. Web image retrieval using self-organizing feature map. *Journal of the American Society for Information Science and Technology*, 52, 2001, 868-875.
Xu, H.; Lancaster, F.W. Redundancy and uniqueness of subject access points in online catalogs. *Library Resources & Technical Services*, 42, 1998, 61-66.
Yang, Y. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1, 1999, 69-90.
Yang, Y. Improving text categorization methods for event tracking. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 65-72. New York, Association for Computing Machinery, 2000.
Yerkey, A.N. Models of index searching and retrieval effectiveness of keyword-in-context indexes. *Journal of the American Society for Information Science*, 24, 1973, 282-286.
Yu, K.-I. et al. Pipelined for speed: the Fast Data Finder system. *Quest*, Winter 1986-1987, 5-19.

Zechner, K. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 199-207. New York, Association for Computing Machinery, 2001.
Zeng, M.L. Metadata elements for object description and representation: a case report from a digitized historical fashion collection project. *Journal of the American Society for Information Science*, 50, 1999, 1193-1208.
Zholkova, A.I. Applying facet analysis methods in abstracting. *Scientific and Technical Information Processing*, 2, 1975, 70-74. (Tradução em inglês de *Nauchno-Tekhnicheskaia Informatsiia*, Seriia 1, número 6, p. 26-28.)
Zhu, B.; Chen, H. Validating a geographical image retrieval system. *Journal of the American Society for Information Science*, 51, 2000, 625-634.
Zich, B. Visualizing digital libraries. In: *Visualizing subject access for 21st century information resources*; ed. by P.A. Cochrane and E.H. Johnson, p. 106-109. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1998.
Zins, C. Models for classifying Internet resources. *Knowledge Organization*, 29, 2002, 20-28.
Zizi, M. Interactive dynamic maps for visualisation and retrieval from hypertext systems. In: *Information retrieval and hypertext*; ed. by M. Agosti and A.F. Smeaton, p. 203-224. Boston, Kluwer, 1996.
Zunde, P.; Dexter, M.E. Factors affecting indexing performance. *Proceedings of the American Society for Information Science*, 6, 1969a, 313-322.
Zunde, P.; Dexter, M.E. Indexing consistency and quality. *American Documentation*, 20, 1969b, 259-267.

ÍNDICE