Indexação automática de textos: uma abordagem otimizada e simples

Jaime Robredo

Resumo

Após uma breve referência às modernas tendências no processamento automático das Informações textuais e às principais abordagens conceituais do processo de indexação automática de textos e mais particularmente à abordagem lingüístico-computacional e à abordagem baseada na eliminação, no texto, das palavras vazias de significado, apresenta-se uma variante desta última abordagem, a *qual permite acelerar, consideravelmente, o processo de escolha dos termos considerados significativos, assim como reduzir de forma importante o volume dos antidicionários de palavras vazias. A abordagem apresentada, já testada com sucesso, integra-se num sistema versátil de indexação automática de textos – o sistema AUTOMINDEX –, o qual, pela* sua *vez, constitui-se num subsistema do sistema BB/DIALOGO. Apresentam-se exemplos de aplicação da nova abordagem do processo de indexação automática da textos que mostram sua flexibilidade.*

Palavras-chave
Recuperação da informação; Indexação automática; Lingüística computacional; Sistema BIB/DIALOGO – AUTOMINDEX.

INTRODUÇÃO

Um amplo estudo comparativo entre indexação manual e automática, baseado na literatura publicada no período de 1970/1983, foi publicado por Vieira[1]. Nesse estudo, a autora destaca a crescente importância que vem adquirindo a indexação automática no decorrer dos últimos anos. Para Lancaster[2], não há dúvida sobre a eliminação progressiva das práticas de indexação manual frente às técnicas de indexação automática. É interessante comparar esta afirmação, já velha de uma década, com as dúvidas formuladas por Courrier[3] oito anos antes, para perceber a velocidade com a qual as coisas têm evoluído *e* continuam a evoluir.

O crescente interesse pelo assunto ficou demonstrado pelo elevado número de comunicações (mais de 20) sobre análise e descrição de conteúdo dos documentos, apresentadas no 44º Congresso da Federação Internacional de Informação e Documentação (FID), realizado em Helsinque, em agosto de 1988[4].

Caberia assinalar, *en passant,* que, como conseqüência da orientação da política de aquisição da maioria das bibliotecas universitárias brasileiras, que privilegia os documentos publicados em língua inglesa, excluindo praticamente a literatura francesa, alemã e russa – pouco procurada pelos pesquisadores anglo-saxônicos – um grande número de publicações da maior relevância torna-se quase inacessível aos pesquisadores brasileiros, tendo estes que realizar um esforço pessoal considerável para obtê-las.

O estudo sobre os processos de indexação automática desenvolveu-se com rapidez como uma das aplicações mais promissoras da inteligência artificial, com vistas à geração automática de bases de dados textuais e à elaboração e a atualização automáticas de dicionários e antidicionários [5,18].

Cabe destacar que, desde um certo tempo, a indexação automática deixou de ser um tema de preocupação exclusiva dos pesquisadores para passar a ocupar, cada vez mais espaço em obras de caráter geral e, inclusive, em livros didáticos[19,23].

O princípio da indexação automática remonta aos últimos anos da década de 50, quando Luhn[24] apresentou o índice KWIC *(key word in context),* no qual as palavras do título que servem de entradas no índice são identificadas automaticamente a partir da eliminação das palavras não significativas, por comparação com uma lista de palavras vazias de significado, estabelecida previamente. A idéia de condensar um texto até reduzi-lo unicamente aos termos realmente significativos é, por outro lado, muito mais antiga, e todos nós a temos aplicado quando redigimos o texto de um telegrama ou de um telex.

Neste trabalho, daremos especial atenção às técnicas de indexação automática que se baseiam na identificação das palavras ou expressões significativas por eliminação, no texto, das palavras vazias. Não trataremos aqui dos métodos de análise estatística da freqüência das palavras ou da atribuição de pesos aos descritores, os quais mais parecem encontrar aplicação na identificação de um elenco de termos integrantes do vocabulário especializado de uma determinada área (que muito podem ajudar na elaboração de dicionários ou de tesauros) do que na escolha efetiva dos descritores representativos do conteúdo informacional de um documento[25,38].

Não trataremos também da indexação automática das perguntas dos usuários, com vistas à montagem das estratégias de busca dos processos de recuperação da informação, aplicação esta que não parece ainda totalmente conclusiva. Entretanto, a especial atenção que vem sendo prestada a esse assunto faz esperar alguns resultados de grande interesse prático nos próximos anos[40,48].

A abordagem lingüístico-computacional, através de processos sucessivos de análise léxica, sintática e semântica das orações do texto, com ajuda de dicionários de palavras vazias e significativas, também procura chegar à identificação dos descritores. Pela sua contribuição teórica merecem uma menção especial os trabalhos de Gardin e sua equipe, os quais culminaram no desenvolvimento do SYNTOL *(syntagmatic organization language),* uma nova e original linguagem documentária com aplicação à identificação de descritores e à recuperação da informação. Mais recentemente, merecem destaque os trabalhos de diversas equipes de pesquisadores europeus[57, 64]. No Brasil, vários pesquisadores vêm se debruçando nos últimos anos sobre o estudo das aplicações do método lingüístico nos processos de indexação e de recuperação da informação[65, 71].

As características diferenciadas das abordagens lingüística e de eliminação das palavras vazias, para identificação dos termos e expressões significativos no texto, não permitem decidir, no momento, sobre a superioridade relativa de uma ou outra no processo de indexação automática[72,74]. Warner enfatiza a necessidade de se realizar um esforço de pesquisa interdisciplinar, abrangendo a lingüística e o processamento de textos em linguagem natural, e aprimorar os estudos sobre recuperação da informação[75].

Nesta rápida revisão dos processos automatizados de análise de textos em linguagem natural, deve ser lembrada *a* amplitude do desenvolvimento das aplicações em sistemas multilingües[76] e, especialmente, dos sistemas de análise e tradução automáticas, que envolvem todos os países da Comunidade Econômica Européia, num esforço exemplar de cooperação e integração[77,81].

Convém destacar que, na atualidade, todos os processos parciais de análise, indexação, organização, armazenagem e recuperação de informações textuais tendem a integrar-se num processo global, consideravelmente mais amplo e abrangente. Esse processo faz parte de uma nova visão da ciência e da engenharia da informação e do conhecimento, como alicerce de um sem número de aplicações práticas que se estendem desde a automação dos escritórios e a edição computadorizada até aos sistemas de informação documentária totalmente automatizados. E isso num ambiente que, graças à incorporação dos sistemas especialistas, nascidos da aplicação da inteligência artificial, facilita cada vez mais o acesso do usuário às informações e aos conhecimentos desejados.

Neste artigo apresentaremos uma abordagem simples e otimizada do processo de indexação automática de textos, a qual pode integrar-se como um módulo específico – evidentemente suscetível de futuras expansões e aprimoramentos – em diversos sistemas aplicativos específicos de complexidade variável.

UMA ABORDAGEM SIMPLES E OTIMIZADA PARA INDEXAÇÃO AUTOMÁTICA DE TEXTOS

Apresenta-se a seguir uma síntese conceituai dos principais elementos que servem de fundamentação aos sucessivos desenvolvimentos do sistema AUTOMINDEX. O algoritmo simplificado do processo de indexação automática encontra-se representado na figura 1, extraída da obra do autor citada anteriormente[82]. No caso concreto do sistema AUTOMINDEX, no lugar de um único dicionário de palavras vazias (ou antidicionário), utilizam-se dois dicionários: I) um dicionário de palavras vazias invariáveis (chamado *wordfixed),* que inclui preposições, conjunções, advérbios etc. e II) um dicionário de raízes de palavras consideradas como não significativas na área de conhecimento considerada (chamado *wordroots).* As figuras 2 e 3 representam, respectivamente, partes dos dicionários de palavras vazias fixas e de raízes das palavras não significativas. As principais vantagens resultantes da utilização desses dois dicionários são as seguintes: I) diminuição do volume total do dicionário; II) diminuição do volume de memória necessário para armazenagem do dicionário; III) diminuição do volume necessário para processamento; IV) aumento da velocidade de processamento.

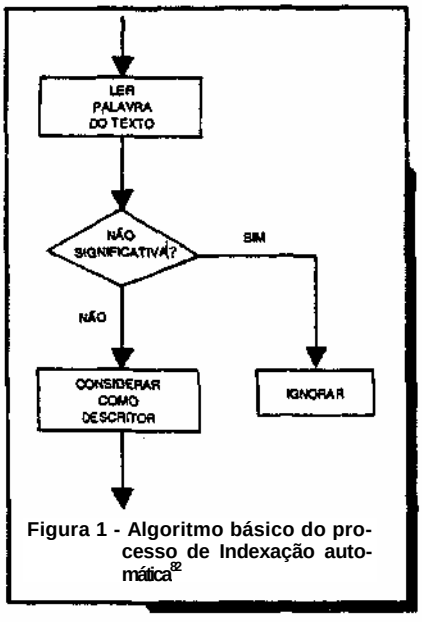

Figura 1 - Algoritmo básico do processo de Indexação automática[82]

A nova abordagem, consistente no uso de dois dicionários de palavras e raízes não significativas, foi utilizada – acreditamos que pela primeira vez – na elaboração automática do índice 1972/1983 da revista *Ciência da Informação*[83], subproduto do estudo comparativo entre indexação automática *e* manual realizado por Vieira[84].

A	CENTRAL	EFICIENTES
ABAIXO	CINCO	EIS
ABRIL	COM	ELA
ACAO	COMO	ELAS
ACERCA	COMPARADA	ELE
ACIMA	CONTRA	ELES
.	.	EM
.	.	EMBORA
.	.	ENFIM
AFORA	DA	ENQUANTO
AGORA	DAÍ	ENTAO
AGOSTO	DALI	ENTRE
AI	DAQUI	.
AINDA	DAS	.
ALEM	DE	.
AMANHA	DELA	ESTA
AMBAS	DELAS	ESTAS
AMBOS	DELE	ESTE
ANO	DELES	ESTES
ANOS	DEMAIS	EXCETO
.	.	.
.	.	.
.	.	.
BOA	DO	FOR
BOAS	DOIS	FORA
BOM	DOS	.
BONS	DOZE	.
.	DUAS	.
.	DURANTE	FRENTE
.	DUZENTOS	.
CA	.	.
CADA	.	.
CEM	.	UM
CENTO	E	UMA
.	.	.
.	.	.
.	.	.

Figura **2- Fragmento do dicionário de palavras vazias de significado ou antidicionário *(wordfixed)***

ABAL	ASPEC	COINCID
ABANDON	ASPIRA	COISA
ABERT	ASSEGUR	COLIG
ABORD	ASSESS	COLOC
ABR	ASSIST	COME
ABSOL	ASSOCIAD	COMPET
.	.	COMPLET
.	.	COMPREE
.	.	COMPRO
ADMIR	BARAT	COMUM
ADMIS	BASEA	CONCED
ADMIT	BASICA	CONCI
ADOT	BASICO	.
ADVER	BASTA	.
AFET	BATE	.
FIRM	BENEF	CONTID
AFLIC	BONI	CONTIN
AFLIG	BRANC	CONTOR
AFRONT	BRILH	CONTRAD
AJUD	BRINC	CONTRAR
.	.	.
.	.	.
.	.	.
ALGUM	CAMINH	CONTRIB
ALGUN	CAMP	.
.	.	.
.	.	.
.	.	CUMPR
APAREC	CHEF	CURIOS
APAREN	CHES	CURT
.	.	.
.	.	.
.	.	.

Figura 3-Fragmento do antidicionário de raízes de palavras não significativas (wordroots)

No processo de indexação automática de textos (títulos e resumos), utilizando-se dos programas delineados pelo autor[85,86], como parte integrante do sistema AUTOMINDEX, utiliza-se como entrada um registro bibliográfico, em formato conforme especificações do sistema BIB/DIALOGO[87, 89]. Numa primeira sub-rotina, são considerados os títulos e resumos dos registros dos documentos e identificadas como "palavras" do texto quaisquer cadeias de caracteres enquadradas entre dois delimitadores previamente estabelecidos (ver figura 4). As sub-rotinas principais varrem o texto e comparam as palavras, primeiramente com os termos do dicionário de palavras vazias fixas e as palavras do texto identificadas no dicionário *wordfixed* são desprezadas. As palavras restantes são varridas de novo e comparadas com as raízes do dicionário *wordroots.* Aquelas palavras cuja raiz é identificada no dicionário são consideradas como não significativas e, conseqüentemente, desprezadas também.

As palavras restantes são consideradas como possíveis descritores. Outra sub-rotina permite sua comparação com um dicionário de raízes significativas, com vistas à sua aceitação como descritores autorizados, em forma normalizada.

Parêntese quadrado (abre)	" ["
Parêntese quadrado (fecha)	"] "
Parêntese (abre)	" ("
Parêntese (fecha)	") "
Ponto	" . "
Dois Pontos	" : "
Vírgula	" , "
Ponto e Vírgula	" ; "
Asterisco	" * "
Apóstrofo	" ' "
Exclamação	" ! "
Interrogação	" ? "
Barra diagonal	" / "
Aspas	" " "
Espaço	" "

Figura 4 – Delimitadores de palavras

Em outra sub-rotina, as "palavras" que não foram identificadas em nenhum dos dicionários, nem como palavras vazias nem como descritores,são gravadas na área do registro reservada aos descritores, para eventual uso em processos posteriores de busca e recuperação, e marcadas corno "candidatos" a desertores para posterior exame e avaliação, com vistas à sua incorporação definitiva ao dicionário de raízes significativas, ou aos dicionários de palavras/raízes vazias de significado.

Finalmente, outras sub-rotinas permitem listar os desertores e candidatos a descritores com suas respectivas freqüências de aparecimento na base de dados, considerando, obviamente, uma vez só os descritores que se repetem no mesmo registro.

A figura 5 representa, com mais detalhe, o algoritmo básico do sistema AUTOMINDEX[90].

Os estudos de freqüência de aparecimento dos descritores a que fizemos referência no início deste trabalho ganham uma importância especial nos processos de elaboração e atualização de dicionários e tesauros, em áreas específicas do conhecimento[91, 92], mas, como já indicamos, esses aspectos não serão considerados neste trabalho.

Cabe dizer, para encerrar esta seção, que uma certa editoração do texto, antes de sua entrada em máquina, permite melhorar, em determinados casos, o nível de qualidade do processo de indexação automática. Assim, a hifenação de vários termos que formam, em conjunto, uma expressão significativa, permite preservar determinados descritores que, de outra forma, não apareceriam ou seriam desdobrados em descritores formados por palavras simples (por exemplo: São-Paulo, Belo-Horizonte, ciência-da-informação, pesquisa-e-desenvolvimento, ciência-e-tecnologia etc.). Nesses casos, uma sub-rotina específica leva a uma nova varredura da expressão hifenada, para identificar as palavras significativas simples, nelas contidas (ciência, informação, pesquisa, desenvolvimento, tecnologia etc.).

É claro que, no momento em que se procede a editoração manual do título e do resumo, pode-se introduzir algum descritor específico ou, ainda, acrescentar algum complemento ao título, como é prática corrente em alguns sistemas internacionais de informação (por exemplo, os sistemas AGRIS e INIS), de forma a enriquecer o significado dos registros documentários, com o qual também pode-se contribuir para melhorar o nível de qualidade de indexação automática.

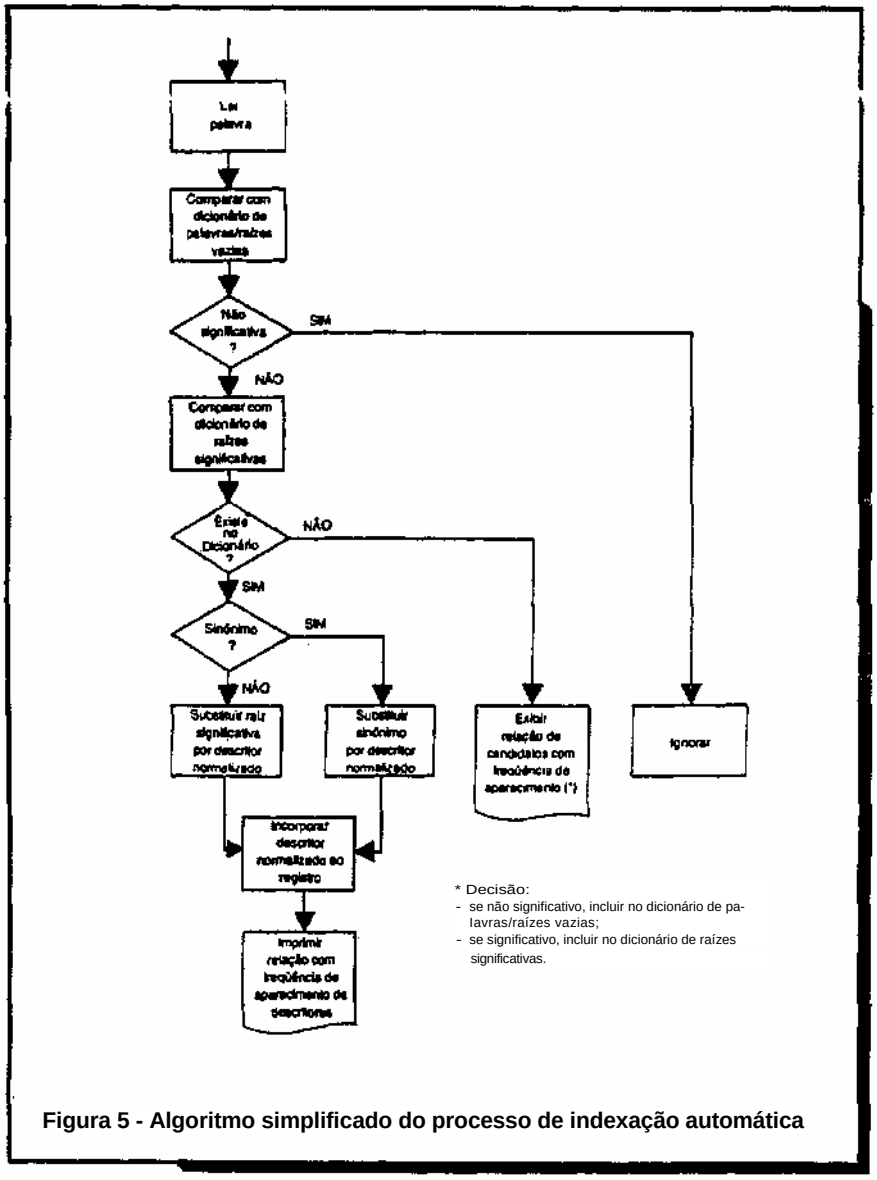

Figura 5 - Algoritmo simplificado do processo de indexação automática

EXEMPLOS DE APLICAÇÃO

Apresentam-se, a seguir, alguns exemplos de aplicação do sistema AUTOMINDEX que mostram a versatilidade da abordagem apresentada, em dois casos bem diferentes:

- indexação automática de registros documentários, a partir dos títulos e dos resumos dos mesmos, para geração de índices temáticos e organização de bases de dados para busca *e* recuperação da informação em linha;

- indexação automática de textos, a partir de documentos do tipo carta, com vistas à organização e arquivo dos documentos num escritório.

A figura 6 permite acompanhar as etapas de preparação do texto e de indexação automática de um registro bibliográfico, a partir do titulo e do resumo. A figura 7 é um fragmento do índice temático dos registros correspondentes aos artigos publicados na revista *Ciência da Informação,* obtido a partir dos descritores extraídos automaticamente a partir dos títulos e dos resumos dos mesmos.

A figura 8 reproduz o texto de uma carta, após editoração, pronto para passar pelo processo de indexação automática. A figura 9 representa os descritores/palavras-chave e/ou candidatos a descritores obtidos automaticamente a partir do texto editado da carta a que se refere a figura 8.

```
a)
CI00119
FREUND, G. E.
UNIVERSIDADE DE SÃO PAULO.
"ANALISE ESTRUTURAL PARA AUMENTAR A EFICIENCIA DE PESQUISAS
ONLINE".
CIENCIA DA INFORMACAO. V.11, N.1, P.19-26. 1982.
PROPOE UMA TECNICA BASEADA NA ANALISE SINTATICA DE TERMOS CO
MO OBJETIVO DE IDENTIFICAR PALAVRAS SEMANTICAMENTE RELACIONA
DAS PARA SOLUCAO DE PROBLEMAS, COMO O DO TRUNCAMENTO ARBITRA
RIO.

b)
ANALISE-ESTRUTURAL PARA AUMENTAR A EFICIENCIA DE PESQUISAS-O
NLINE.
PROPOE UMA TECNICA BASEADA NA ANALISE-SINTATICA DE TERMOS CO
M O OBJETIVO DE IDENTIFICAR PALAVRAS-SEMANTICAMENTE-RELACION
ADAS PARA SOLUCAO DE PROBLEMAS, COMO O DE TRUNCAMENTO-ARBITR
ARIO.

c)
ANALISE-ESTRUTURAL      PALAVRAS-SEMANTICAMENTE-RELACIONADAS
ANALISE                 palavras (PALAVRAS)
estrutural (ESTRUTURA)  semanticamente (SEMANTICA)
EFICIENCIA              TRUNCAMENTO-ARBITRARIO
PESQUISAS-ONLINE        TRUNCAMENTO
pesquisas (PESQUISA)
ONLINE
TECNICA
ANALISE-SINTATICA
sintatica (SINTAXE)
termos (TERMO)
```

Figura 6 - Processo de indexação automática a partir dos títulos e resumos dos registros documentários

```
UNIVERSIDADE-FEDERAL-DA-BAHIA
ESCOLA-DE-BIBLIOTECONOMIA-E-DOCUMENTACAO
SALVADOR, 03-ABR-89

ILMO. SR.
PROF. JAIME-ROBREDO
BRASILIA, DF

PREZADO PROFESSOR,
DE CONFORMIDADE COM OS ENTENDIMENTOS POR TELEFONE, MANTIDOS COM
V.SA., CONFIRMAMOS O PERIODO DE 03-07 JUL PARA A REALIZACAO DO
CURSO DE INTRODUCAO AOS PROCESSOS DE INDEXACAO-AUTOMATICA DE TE
XTOS. INFORMAMOS, OUTROSSIM, QUE O REFERIDO CURSO JÁ SE ENCONTR
A APROVADO PELA COORDENACAO-DE-EXTENSAO DA UFBA, DE CONFORMIDAD
E COM O PROGRAMA E CARGA-HORARIA DE 40 HORAS, ESTABELECIDOS POR
V.SA.
ESTAMOS ANCAMINHANDO, NA PRESENTE DATA, OFICIO DO DR. MARCOS-FO
RMIGA, DO INEP, VISANDO A CONFIRMACAO DA PASSAGEM-AEREA BRASILI
A-SALVADOR-BRASILIA, A SER UTILIZADA POR V.SA.  PARA REALIZACAO
DO MESMO CURSO.

ATENCIOSAMENTE.
MARGARIDA-PINTO-DE-OLIVEIRA
COODENADORA-DO-CURSO
```

Figura 8 - Texto de uma carta, após editoração

FONTE	CI00050
FONTE	CI00126
FONTES	CI00064
FONTES-DE-INFORMAÇÃO	CI00047
FORMAÇÃO	CI00099
FORMAÇÃO-PROFISSIONAL	CI00019
FORMATO	CI00011
FÓRMULA-DE-TRANSIÇÃO	CI00024
FORNECEDORES-DA-INFORMAÇÃO	CI00144
FORNECIMENTO-DE-LIVROS	CI00005
FORTALECIMENTO	CI00138
FRANCA	CI00061
FRASES	CI00025
FRENTE-DE-PESQUISA	CI00016
FRENTE-DE-PESQUISA	CI00044
FRENTE-DE-PESQUISA	CI00094
FREQUENCIA	CI00029
FREQUENCIA	CI00032
FUTUROLOGIA	CI00019
GATEKEEPERS	CI00109
GEOLOGICA	CI00017
GEORGES-ANDERLA	CI00071
GERAÇÃO	CI00114

Figura 7 - Fragmento do índice temático impresso, obtido mediante o processo de indexação automática, a partir do título e do resumo dos artigos publicados na revista *Ciência da Informação*[83]

```
UNIVERSIDADE-FEDERAL-DA-BAHIA
UNIVERSIDADE
FEDERAL
BAHIA
ESCOLA-DE-BIBLIOTECONOMIA-E-DOCUMENTACAO
BIBLIOTECONOMIA
DOCUMENTACAO
SALVADOR
03-ABR-89
JAIME-ROBREDO
BRASILIA
DF
entendimento (ENTENDIMENTO)
TELEFONE
confirmamos (CONFIRMACAO)
03-07-JUL
CURSO
INTRODUCAO
PROCESSOS
INDEXACAO-AUTOMATICA
INDEXACAO
automatica (AUTOMACAO)
textos (TEXTO)
informamos (INFORMACAO)
aprovado (APROVACAO)
COORDENACAO-DE-EXTENSAO
COORDENACAO
EXTENSAO
UFBA
PROGRAMA
CARGA-HORARIA
40-HORAS
encaminhamos (ENCAMINHAMENTO)
DATA
OFICIO
MARCOS-FORMIGA
INEP
PASSAGEM-AEREA
BRASILIA-SALVADOR-BRASILIA
MARGARIDA-PINTO-DE-OLIVEIRA
COORDENADORA DO CURSO
```

Figura 9 - Descritores/palavras-chave e/ou candidatos extraídos do texto da carta a que se refere a figura 8

CONCLUSÃO

Nos dias atuais parece difícil continuar a se duvidar da superioriedade da qualidade da indexação automática com relação à indexação manual, quando a abordagem conceitual que serve de fundamento às rotinas dos programas de análise e processamento dos textos seguem determinadas regras claramente estabelecidas.

Dentre as diversas abordagens que norteiam os vários métodos de identificação e escolha dos desertores ou das palavras-chave, a partir dos textos escritos em linguagem natural, a abordagem de seleção por exclusão permite desenvolver mecanismos simples e eficazes, com grande versatilidade de aplicação.

A associação dos avanços da inteligência artificial aos métodos de indexação automática deverá abrir, nos anos vindouros, perspectivas do maior interesse às aplicações práticas das técnicas de análise e indexação automáticas de textos.

No Brasil, contrasta, no momento, o limitado número de pesquisas realizadas ou em curso, nessa área, com o esforço gigantesco que está sendo realizado nos últimos anos, nos países industrializados.

Um maior esforço de pesquisa nessa área e maior comunicação entre os pesquisadores poderiam contribuir para reduzir rapidamente a defasagem observada, abrindo interessantes possibilidades de desenvolvimento a novos sistemas aplicativos.

REFERÊNCIAS BIBLIOGRÁFICAS E NOTAS

1. VIEIRA, Simone Bastos. Indexação automática e manual: revisão da literatura. *Ciência da Informação,* v. 17, n. 1, p. 43-57, jan./jul 1988.

2. LANCASTER, F.W.Trends in subject indexing from 1957 to 2000. In: CONGRESS, 39. Edinburgh, 25-28 September 1978. New trends in documentation and information. London: Aslib, 1980. p. 223-233.

3. COURRIER, Yves. L'indexation automatique: état de Ia question et perspectives d'avenir. *Documentation et bibliothèques,* v. 23, n. 2, p. 59-72, juin 1977.

4. FID Conference and Congress, 44. Helsinki, 28 August 1 September 1988 (Participants edition). Helsinki: International Federation for Information and Documentation; Finish Society for Information Services. 1988. Part 1. p. 146-358.

5. WILHS, Yorick. Natural Language understanding systems within the R. I. Paradigm: a survey. *American Journal of Computational Linguistics,* n. 1,1976. (Microfiche n. 40).

6. WAHLSTER, W. *Natuerlichsprachliche Systeme -Eine Einfuehrung in die sprachorientierte KI-Forschung.* Bericht GEN-10, Forschungstelle fuer informationswissenschaft und Kuenstliche Inteligenz. Universitaet Hamburg, 1982. Citado por KNORZ, Gerhard *Automatisches Indexiaren als Erkennen abstrakter Objekte.* Tuebingen: Niemeyer. 1983 (Sprache und Information: Bd.8).

7. _____ *Die Represaentation von vagen Wissen in natuerlichsprachlichen Systemen der Kuenstlichen Intelligenz.* Diplomarbeit. (If I ─ HH B ─ 38/77). Instituet fuer Informatik, Universitaet Hamburg. Juli 1977.

8. NOEL, J., MOULIN, A., JANSENS WHITE, J. B. Univ. Liege. Service Langue Anglaise Moderne. *Cahiers de la documentation,* v. 40, n. 2, p. 68-79, 1980.

9. TUFFERY, Michael. *Système documentaire, base de données textuelles:* le projet Etoile. Doct. 3e cycle. Informatique. Toulouse 3. 1984.

10. GOEFENSTETTE, Gregory. *Traitements linguistiques orientés vers Ia documentation automatique.* Doct. 3e cycle. Electronique. Paris 11. 1983.

11. AIT HAMLAT, Aleila. *Applixations des méthodes statistiques à l'indexation automatique de documents.* Doct. 3e cycle. Math. Paris 6. 1983.

12. KERKOUBA, Dalila. *Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles:* application à un corps technique. Doct, 3e cycle. Informatique. INP Grenoble.1984,

13. SEGUIN, Gerard. *Génération automatique du vocabulaire representatif d'un domaine:* essais d'indexation automatique. Doct. 3e cycle. S.C. Lyon 1. 1977. ≠ 626.

14. LAINE, Sylvie. *Extraction et selection de descripteurs complexes dans un ensemble de textes pour leur indexation automatique.* Doct. ingenieur. Math. Lyon 1.1982.

15. MERLE, Alain. *Un analyseur pré-syntaxique pour la levée des ambiguités dans des documents écrits en langue naturelle:* application à l'indexation automatique. Doct. ingenieur. Informatique. INP Grenoble. 1982.

16. CHARTRON, Ghislaine.*Analyse des corpus de données textuelles:* sondage de flux d'informations. Thèse de nouveau doctorat en traitement de l'information. Paris 7. Juin 1988. 273 p.

17. MICHELET, Bertrand.*L'analyse des associations.* Thèse de nouveau doctorat en, traitement de l'information. Paris 7. Oct. 1988.409p.

18. CHARTRON, Ghislaine. Lexicon management tools for large textual databases: the Lexinet system. *Jornal of Information Science,* n. 15. p. 2-10,1989.

19. COYAUD, Maurice, SIOT-DECAUVILLE, Nelly. *L'analyse automatique des documents.* Paris: Mouton. 1967.147 p.

20. CLEVELAND, Donald B., CLEVELAND, Ana D. *Introduction to indexing and abstracting.* Littleton: Libraries Unlimited. 1983.

21. COLL-VINENT, Robert. *Información y poder:* el futuro de las bases de datos documentales. Barcelona: Herder. 1988. 295 p.

22. ROBREDO, Jaime, CUNHA, Murilo Bastos da. *Documentação de hoje e de amanhã:* uma abordagem informatizada da bibliteconomia e da ciência da informação. 2. ed. Brasília: ed. autor. 1986. 400 p.

23. GUINCHAT, Claire, SKOURI, Yolande. *Guide pratique des techniques documentaires.* Traitement et gestion des documents, v. 1, Paris: EDICEF, 1989. p. 243-245.

24. LUHN, H. P.*Key-word-in-context index for technical literature (KWIC index).* IBM Advanced System Development Division Report. RC-12.1959.

25. DIONNE, Guy. PRECIS I: Preserved Context Indexing System. *Documentation et Bibliothèques,* v. 21, n. 1, p. 9-12, Mars 1975.

26. AUSTIN, Derek. *PRECIS. a manual of concept analysis in subject indexing.* London; BNB, 1974.

27. SPARK-JONES, Karen, KAY, Martin. *Linguistics and Information science*. New York: Academic Press. 1973.

28. BOOKSTEIN, A., SWANSON, D. R. Probabilistic models for automatic indexing. *Journal of the American Socieiy for Information Science,* n. 25, p. 312-318, 1975.

29. CAROLL, J.M., ROELOFFS, R. Computer selection of keywords using word-frequency analysis. *American Documentation,* n. 25,1969.

30. HARTER, S.D. A probabilistic approach to automatic keyword Indexing. Part I. On the distribution of speciality words in a technical literature. Part II. An algorithm for probabilistic indexing. *Journal of the American Society for Information Science,* n. 26, p. 197-206, p. 280-289.1975.

31. KRAGENINGS, R. *Statistische Relationen zwischen Textwoerten und Deskriptoren.* (ZMD-R-25), Beuth Verlag. 1974: 33. Citado por KNORZ, Gerhard; vide supra (6).

32. SALTON, G. *Experiments in automatic thesaurus construction for information retrieval.* Information Processing 71. Amsterdam: North Holland. 1972. p. 115-123.

33._____ .On the specification of term values in automatic indexing. *Journal of Documentation, v.* 29, n. 4, p. 351-372. 1973.

34. _____ . Binary and weighted retrieval. In: ODELY, R.N., ROBERTSOD, S.E., VAN RIJSBERGEN, C.J., WILLIAMS, P. (coord.). *Information Retrieval Research.* London: Butterworth, 1981. p. 9-22.

35. _____ . WONG. A.; YU, C.T. Automatic indexing using term discrimination and term precision measurement. *Information Processing and Management,* n. 12, p. 43-56,1976.

36. PROJEKT Wai. *Woerterbuchentwicklung fuer automatisches Indexing.* Sachbericht 1980. (DV 80-3). Technische Hochschule Darmstadt, Fachbereich Informatik. FG. DVS II, 1980.17. Citado por KNORZ, Gerhard; vide supra 6.

37. KWOK, K. L, A probabilistic theory of Indexing using authorprovided relevance information. American Society for information Science Anual Meeting. 48. *Proceedings.* v. 22. White Plains: Knowledge Industry. 1983. p. 59-63.

38. DEVADASON, Francis Jawahar. Computer generation of different types of subject entries based on deep structure of subject indexing languages: deep structure indexing system. American Society for Information Science Annual Meeting. 48. *Proceedings.* v. 22. White Plains: Knowledge Industry, 1985. p. 88-98.

39. FUJITA, Mariângela L. Sistema de indexação PRECIS l. PRECIS ─ perspectiva histórica e técnica de desenvolvimento e aplicação. *Revista Brasileira de Biblioteconomia e Documentação,* v. 21, n. 1/2, p. 21-45, jan./jun. 1988.

40. SMMITH, L. Artificial intelligence in Information retrieval systems. *Information Processing and Management,* n. 12, p. 189-222, 1979.

41. FOX, Edward A. Analysis and retrieval of composite documents. American Society for Information Science Annual Meeting. 48. *Proceedings.* V. 22. White Plains: Knowledge Industry. 1985. p. 54-58.

42. CRAWFORO, R. G., Becker, H. S. *Toward the development of interfaces for untrained users. American Society for Information Science Annual Meeting. 48. Proceedings.* V. 22. White Plains: Knowledge Industry. p. 236-239.

43. MARCUS, Richard S. Development and testing of expert systems for retrieval assistance. American Society for Information Science Annual Meeting. 48. *Proceedings.* V. 22. White Plains: Knowledge Industry. 1985. p. 289-292.

44. METZLER, Douglas P., NOREAULT, Terry, HAAS, Douglas F., COSIC, Cynthia. An expert system approach to natural language processing. American Society for Information Science Annual Meeting. 48. *Proceedings.* V. 22. White Plains: Knowledge Industry, 1985. p. 301-307.

45. CHAMIS, Alice Yanosko. The usefulness of switching vocubularies for on-line databases. American Society for Information Science Annual Meeting. 48. *Proceedings.* v. 22. White Plains: Knowledge Industry. 1985. p. 311-314.

46. ALMEIDA, Carlos Henrique Marcondes de. Uma interface inteligente para usuários casuais de sistemas de recuperação de informações remotos. Congresso Nacional de Informática. 22./Seminário Nacional de Biblioteconomia e Informática 7. Rio de Janeiro. 27-31 de agosto de 1990.

47. CUNHA, Isabel Maria R. Ferin. Perspectivas da análise documentária em inteligência artificial. Congresso Nacional de Informática. 22./Seminário Nacional de Biblioteconomia e Informática. 7. Rio de Janeiro. 27-31 de agosto de 1990.

48. ROY, Galles. Intelligent user interface for information retrieval. Congresso Nacional de Informática. 22. Congresso Internacional de Informática. 3. Rio de Janeiro. 27-31 de agosto de 1990.

49. GARDIN, J. C. Document analysis and linguistic theory. *Journal of Documentation,* v. 29, n. 2, p. 137-168, June 1973.

50. _____ *Les analyses du discours.* Neuchátel: Delachaux et Niestlé. 1974.

51. ALOUCHE, F. et al. *Economie génèrale d'une chaine documentaire mécanisée.* Paris: Gauthier-Villars. 1967.

52. BELY, N. et al. *Procédures d'analyse sémantique appliquées à Ia documentation scientifique.* Paris. Gauthier-Villars. 1970.

53. COYAUD, M. *Linguistique et documentation.* Paris: Larousse. 1972.

54. LEVY, F. General economy of a mechanical documentation chain. *Unesco Bull. Lib. v.* 25, n. 5, p. 232, Sept./Oct. 1967.

55. CROS, R. C., GARDIN, J. C., LÉVY, F. *L'automatisation des recherches documentaires:* un modèlo général, le SYNTOL. Paris: Gauthier-Villars. 1964.

56. COURRIER, Yves. SYNTOL. In: KENT, Allen, LANCOUR, Harold, DAILY, Jay E. (ed.) *Encyclopedia of Library and Information Sciences.* New York: Marcel Decker. 1976. p. 375-380.

57. ZIMMERMANN, H. Ansaetze einer realistischen automatischen Indexierung unter Verwendung linguistischen Verfahren. In: KUEHLEN, R. (coord.) *Datenbasen, Datenbanken, Netzwerke.* Praxis der Information Retrieval, Bd. 1. Verlag Dokumentation. 1979. p. 311-338.

58. WOLF, J. J., WOODS, W. A. the WHIM speech understanding system. In: LED, W. A. (coord.). *Trends in speech recognition.* Englewood Cliffs, N. J.: Prentice-Hall, 1980. Chap. 14.

59. ROSTEK, L. Verstehen und Verdichten von Texten-Represaentation und Strategie bei der Verarbeitung in einen Modell fuer ein Referenz-Retrievalsystem, basierend auf einem Netz von Begriffsbezeichnungen. In: BATORI, l., KRAUSE, J., LUTZ, H. D. (coord.). *Linguistische Datenverarbeitung ─ Versuch einer Standortbestimmung in Umfeld von Informationlinguistik und kuenstlicher Inteligenz.* Tuebingen: Niemeyer. 1982. p. 127-136 (Sprache und Information Bd. 4).

60. ROLLINGER, C.R., Readtime ─ Inferenzen fuer semantische Netze. In: ROLLINGER, C. R., SCHNEIDER, H. J. (coord.). *Interenzen in natuerlichsprachlichen Systemen in der K. I.* Berlin: Einhom, 1980. p. 115-150.

61. MORIK, K. Differenzstudie zu frueheren sprachverarbeitenden Systemen der Bundersrepublic Deutschland. Forschunstelle fuer Informationwissenschaft und kuenstliche Intelligenz. Universitaet Hamburg. Report ANS-6. 1982.

62. MAAS, H. D. Repraesentation und Strategie bei der automatischen Analyse mit SUSY. In: BATORI, I., KRAUSE, J., LUTZ, H. D. (coord.). *Linguistische Datenverarbeitung ─ Versuch einer Standortbestimmung im Umfeld von Informationslinguistik und kuenstlicher Intelligenz.* Tuebingen: Niemayer. 1982. p. 105-110 (Sprache und Information Bd. 4).

63. MARBURGER, H. Ueberlegungen zur Entwicklung eines deutschsprachingen Zugangssystems zu formatierten Massendaten. Forschungsstelle fuer Informationswissenschaft und kuenstliche Intelligenz. Universitaet Hamburg, Memo ANS-9. 1982.

64. HABEL, Chr. Prableme und Konzepte der maschinellen Sprachverarbeitung aus der Sicht der kuenstlichen Intelligenz und Cognitive Sciece. In: BATORI, l., KRAUSE, J., LUTZ, H. D. (coord.). *Linguistische Datenverarbeitung — Versuche einer Standortbestimmung im Umfeld von Informationslinguistik und kuenstlicher Intelligenz.* Tuebingen: Niemeyer, 1982. p. 39-56. (Sprache und Information Bd. 4).

65. ANDREEWSKI, Alexandre, RUAS, Vitoriano. *Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa.* Rio de Janeiro: Pontifícia Universidade Católica. 1982. 31 p. (Monografias em Ciência da Computação, n. 3/82).

66. CINTRA, Ana Maria Marques. Elementos de lingüística para estudos de indexação. *Ciência da Informação,* v. 12, n. 1, p. 5-22, jan./jun. 1983.

67. BARANOW, Ulf Gregor. Perspectivas na contribuição da lingüística e de áreas afins à ciência da informação. *Ciência da Informação,* v. 12, n. 1, p. 61-69, jan./jun. 1983.

68. ANDREEWSKI, Alexandre, RUAS, Vitoriano. Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa. *Ciência da Informação,* v. 12, n. 1, p. 61-69, jan./jun. 1983.

69. HALLER, Johann. Processamento de textos em linguagem natural. In: *Congresso Nacional de Informática.* 15. Rio de Janeiro, Out. 1982. (Trabalhos apresentados). 9 p.

70. _____Análise automática de textos em sistemas de informação. *Revista de Biblioteconomia de Brasília,* v. 11, n. 1, p. 105-113, jan./jun. 1983.

71. NAVARRO, Sandrelei. Interface entre lingüística e indexação. *Revista Brasileira de Biblioteconomia e Documentação,* v. 21, n. 1/2, p. 46-62, jan./jun. 1988.

72. BORKO, Harold, BERNIER, Charles, L. *Indexing concepts and methods.* New York: Academic Press, 1978. p. 117.

73. ROBREDO, Jaime, CUNHA, Murilo Bastos da. op. cit p. 256.

74. NIEMISTO, J., JAEPPINEN. H. Some linguistic problems for automatic indexing caused by inflections and compounds. In: FID Conference and Congress, 44., Helsinki, 28 August-1 September 1988.

(Participants edition). Helsinki: FID; FSIS, 1988, Part 1. p. 158-163.

75. WARNER, Amy J. Linguistic theories for information retrieval. In: FID Conference and Congress, 44. Helsinkl, 28 August-1 September 1988. (Participants edition), Helsinki: FID; FSIS, 1988. Part 1. p. 146-157.

76. DUBOIS, C. P. R. Multilingual information systems: some criteria for the choice of specific techniques. *Journal of Informations Science*, n. 1, p. 5-12,1979.

77. HALLER, Johann. Das EUROTRA-Projekt─ Stand 1987 und Ausblick. *Sprache und Datenverarbeitung,* v. 11, n. 1, p. 5-7. 1987.

78. _____ . Anwendung linguistischer Forschungergebnisse in der Maschinellen-Uebersetzung: Die Diskusion der Interface-Struktur (IS) in EUROTRA. *Sprache und Datenverarbeitung,* v. 11, n. 1, p. 8-14. 1987.

79. MAAS, Heinz- Dieter. The dictionary in the EUROTRA engineering framework. *Sprache und Datenverarbeitung, v.* 11, n. 1, p. 15-21.1987.

80. GOESER, Sebastian. Zur Konzeption eines Parsers fuer realistische Textanalyse. *Sprache und Datenverarbeitung,* v. 11, n. 1, p. 43-48.1987.

81. HALLER, Johann. *L'application de théories linguistiques dans Ia mise au point d'un système de traduction automatique (avec référence particulière au Projet EUROTRA).* Doct. 3. cycle. Paris. 1989. 83 p.

82. ROBREDO, Jaime, CUNHA, Murilo Bastos da, op. cit. p. 255.

83. CIÊNCIA DA INFORMAÇÃO: Índice 1972-1983. Brasília: CNPq/IBICT; UnB/BIB. 1985. 79 p.

84. VIEIRA, Simone B. *Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação.* Brasília: Universidade de Brasília/Departamento de Biblioteconomia, 1984. (Dissertação de mestrado). V. também: VIEIRA, Simone Bastos. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação*. Revista de Biblioteconomia de Brasília,* v. 16, n. 1, p. 83-94. jan./jun. 1988.

85. ROBREDO, Jaime, FERREIRA, José Adalberto de P. Conceituação de um programa para Indexação automática de textos. *Revista de Biblioteconomia de Brasília*, v. 8, n. 2, p. 254-263, jul./dez. 1980.

86. ROBREDO, Jaime. A indexação automática: o presente já entrou no futuro. In: MACHADO, Ubaldino Dantas (ed.) Estudos avançados em ciência da informação. Brasília. Associação dos Bibliotecários do Distrito Federal, 1982.

87. _____ . Dois novos sistemas com computador para o processamento completo da informação documentária. *Revista de Biblioteconomia de Brasília,* v. 10, n. 1, p. 69-72, jan./fev. 1982.

88. BIB/Batch & BIB/Dialog. In: KEREN, C. SERED, G, (ed.). International inventory of software packages in information field. Paris: Unesco, 1983. p. 119-124. (PGI/UNISIST. PGI/83/WS/28).

89. ROBREDO, Jaime. Uma experiência de aplicação do computador no ensino da biblioteconomia e ciência da informação. *Revista de Biblioteconomia de Brasília,* v. 12, a 1, p. 11-24, jan./jun.1984.

90. _____ . A indexação automática como mecanismo básico no processo de transferência da informação. In: Congresso Latino americano de Biblioteconomia e Documentação, 1. Salvador, 21-28 Set. 1980. (Trabalhos apresentados). 19 p.

91. ROBREDO, Jaime. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos adequados de controle terminológico. *Ciência da Informação,* v. 11, *n.* 1, p. 3-18, jan./jun. 1982.

92. MOREIRA, José de Albuquerque. *Elaboração de um dicionário de raízes significativas na literatura brasileira de biblioteconomia, documentação e ciência da informação.* Brasília: Universidade de Brasília/Departamento de Biblioteconomia. 1986. (Dissertação de mestrado).

Artigo aceito para publicação em 19 de abril de 1991.

Jaime Robredo

Doutor em Ciências, professor titular do Departamento de Biblioteconomia da Faculdade de Ciências Sociais Aplicadas da Universidade de Brasília.

Automatic text indexing: an improved and simple approach

Abstract

Following a short reference to the modern trends in the automatic processing of textual information, as well as to the most relevant conceptual approaches to the processes of automatic text indexing and, more specifically, the computational-linguistic approach and that one based on the elimination in the text of non-significant words, a modification of this last approach is described, which makes it possible to accelerate significantly the process of identification of the terms considered significant, reducing, at the same time, the volume of the anti-dictionnaries of stopwords. The new approach, which has been successfully tested and applied, is a component of a versatile system of automatic indexing of texts ─ the AUTOMINDEX System ─, this being, in turn, an integrating part of the System BIB/DIALOGO. Examples of application of the new approach of automatic text indexing are given, which demonstrate its versatility and flexibility.

Key words
Information retrieval; Automatic indexing; Computational Linguistics; BIB/DIALOG, AUTOMINDEX Systems.