Metodologia de Pesquisa para Ciência da Computação

Profa. M. Cristina
ICMC-USP

SCC5921 – Metodologia de Pesquisa em Visualização e Imagens

Understanding the causes of crime is a longstanding issue in researcher's agenda. While it is a hard task to extract causality from data, several linear models have been proposed to predict crime through the existing correlations between crime and urban metrics. However, because of non-Gaussian distributions and multicollinearity in urban indicators, it is common to find controversial conclusions about the influence of some urban indicators on crime. Machine learning ensemble-based algorithms can handle well such problems. Here, we use a random forest regressor to predict crime and quantify the influence of urban indicators on homicides. Our approach can have up to 97% of accuracy on crime prediction, and the importance of urban indicators is ranked and clustered in groups of equal influence, which are robust under slightly changes in the data sample analyzed. Our results determine the rank of importance of urban indicators to predict crime, unveiling that unemployment and illiteracy are the most important variables for describing homicides in Brazilian cities. We further believe that our approach helps in producing more robust conclusions regarding the effects of urban indicators on crime, having potential applications for guiding public policies for crime control.

Document clustering is a necessary step in various analytical and automated activities. When guided by the user, algorithms are tailored to imprint a perspective on the clustering process that reflects the user's understanding of the dataset. More than just allow for customized adjustment of the clusters, a visual analytics approach will provide tools for the user to draw new insights on the collection. While contributing his or her perspective, the user will also acquire a deeper understanding of the data set. To that effect, we propose a novel visual analytics system for interactive document clustering. We built our system on top of clustering algorithms that can adapt to user's feedback. In the proposed system, initial clustering is created based on the user-defined number of clusters and the selected clustering algorithm. A set of coordinated visualizations allow the examination of the dataset and the results of the clustering. The visualization provides the user with the highlights of individual documents and understanding of the evolution of documents over the time period to which they relate. The users then interact with the process by means of changing key-terms that drive the process according to their knowledge of the documents domain. In key-term-based interaction, the user assigns a set of key-terms to each target cluster to guide the clustering algorithm. We have improved that process with a novel algorithm for choosing proper seeds for the clustering. Results demonstrate that not only the system has improved considerably its precision, but also its effectiveness in the document-based decision making. A set of quantitative experiments and a user study have been conducted to show the advantages of the approach for document analytics based on clustering. We performed and reported on the use of the framework in a real decision-making scenario that relates users discussion by email to decision making in improving patient care. Results show that the framework is useful even for more complex data sets such as email conversations.

We present a novel architecture, the “stacked what-where auto-encoders” (SWWAE),which integrates discriminative and generative pathways and provides a unified approach to supervised, semi-supervised and unsupervised learning without relying on sampling during training. An instantiation of SWWAE uses a convolutional net (Convnet) (LeCun et al. (1998)) to encode the input, and employs a deconvolutional net (Deconvnet) (Zeiler et al. (2010)) to produce the reconstruction. The objective function includes reconstruction terms that induce the hidden states in the Deconvnet to be similar to those of the Convnet. Each pooling layer produces two sets of variables: the “what” which are fed to the next layer, and its complementary variable “where” that are fed to the corresponding layer in the generative decoder.

Face recognition is a long-standing challenge in the field of Artificial Intelligence (AI). The goal is to create systems that detect, recognize, verify and understand characteristics of human faces. There are significant technical hurdles in making these systems accurate, particularly in unconstrained settings, due to confounding factors related to pose, resolution, illumination, occlusion and viewpoint. However, with recent advances in neural networks, face recognition has achieved unprecedented accuracy, built largely on data-driven deep learning methods. While this is encouraging, a critical aspect limiting face recognition performance in practice is intrinsic facial diversity. Every face is different. Every face reflects something unique about us. Aspects of our heritage – including race, ethnicity, culture, geography – and our individual identity – age, gender and visible forms of self-expression – are reflected in our faces. Faces are personal. We expect face recognition to work accurately for each of us. Performance should not vary for different individuals or different populations. As we rely on data-driven methods to create face recognition technology, we need to answer a fundamental question: does the training data for these systems fairly represent the distribution of faces we see in the world? At the heart of this core question are deeper scientific questions about how to measure facial diversity, what features capture intrinsic facial variation and how to evaluate coverage and balance for face image data sets. Towards the goal of answering these questions, Diversity in Faces (DiF) provides a new data set of annotations of one million publicly available face images for advancing the study of facial diversity. The annotations are generated using ten facial coding schemes that provide human-interpretable quantitative measures of intrinsic facial features. We believe that making these descriptors available will encourage deeper research on this important topic and accelerate efforts towards creating more fair and accurate face recognition systems.

Deep learning is quickly becoming the standard technique for image classification. The main problem facing the automatic identification of plant diseases using this strategy is the lack of image databases capable of representing the wide variety of conditions and symptom characteristics found in practice. Data augmentation techniques decrease the impact of this problem, but those cannot reproduce most of the practical diversity. This paper explores the use of individual lesions and spots for the task, rather than considering the entire leaf. Since each region has its own characteristics, the variability of the data is increased without the need for additional images. This also allows the identification of multiple diseases affecting the same leaf. On the other hand, suitable symptom segmentation still needs to be done manually, preventing full automation. The accuracies obtained using this approach were, in average, 12% higher than those achieved using the original images. Additionally, no crop had accuracies below 75%, even when as many as 10 diseases were considered. Although the database does not cover the entire range of practical possibilities, these results indicate that, as long as enough data is available, deep learning techniques are effective for plant disease detection and recognition.

The Abstract

Informative

Contains all the relevant information of the paper

x

Descriptive

Describes only the nature/purpose of the study

- ***Setting/Background***
- ***Gap***
- ***Purpose***
- ***Methodology***
- ***Main Results***
- ***Conclusion***

Sound emissions by ships and boats can strongly impact marine life, with potential to affect communications, breeding and prey and predator relationships. Automatic detection of boat signatures in underwater audio recordings is thus an important task. Automated solutions are particularly relevant for monitoring preservation areas where the presence of watercrafts is usually regulated. The task is particularly challenging because it requires distinguishing multiple overlapping acoustic events in typically noisy audio recordings. In this paper we introduce an algorithm for boat and ship detection which computes an acoustic signature that captures the variance in the frequency amplitudes observed over the duration of the signal. We evaluated the algorithm on a database of underwater recordings collected at two conservation areas in the State of São Paulo, Brazil, with very good results, and also compared it with an existing solution. Besides being effective, the algorithm requires limited user input and no parameter fine tuning to handle diverse situations. It thus provides a solution to automate the detection of vessels, with potential applications for monitoring marine preservation areas.

Reis et al. Automatic Detection of Vessel Signatures in Audio Recordings with Spectral Amplitude Variation Signature. *Methods in Ecology and Evolution 10(9)*, 2019.

1) Context: Sound emissions by ships and boats can strongly impact marine life, with potential to affect communications, breeding and prey and predator relationships. Automatic detection of boat signatures in underwater audio recordings is thus an important task. Automated solutions are particularly relevant for monitoring preservation areas where the presence of watercrafts is usually regulated.

2) GAP: The task is particularly challenging because it requires distinguishing multiple overlapping acoustic events in typically noisy audio recordings.

3) Purpose: In this paper we introduce an algorithm for boat and ship detection

4) Methodology: which computes an acoustic signature that captures the variance in the frequency amplitudes observed over the duration of the signal.

5) Results: We evaluated the algorithm on a database of underwater recordings collected at two conservation areas in the State of São Paulo, Brazil, with very good results, and also compared it with an existing solution.

6) Conclusions: Besides being effective, the algorithm requires limited user input and no parameter fine tuning to handle diverse situations. It thus provides a solution to automate the detection of vessels, with potential applications for monitoring marine preservation areas.

Many real-world networks display hidden community structures with important potential implications in their dynamics. Many algorithms highly relevant to network analysis have been introduced to unveil community structures. Accurate assessment and comparison of alternative solutions are typically approached by benchmarking the target algorithm(s) on a set of diverse networks that exhibit a broad range of controlled features, ensuring the assessment contemplates multiple representative properties. Tools have been developed to synthesize bipartite networks, but none of the previous solutions address the issue of generating networks with overlapping community structures. This is the motivation for the BNOC tool introduced in this paper. It allows synthesizing bipartite networks that mimic a wide range of features from real-world networks, including overlapping community structures. Multiple parameters ensure flexibility in controlling the scale and topological properties of the networks and embedded communities. BNOC's applicability is illustrated assessing and comparing two popular overlapping community detection algorithms on bipartite networks, namely HLC and OSLOM. Results reveal interesting features of the algorithms in this scenario and confirm the relevant role played by a suitable benchmarking tool. Finally, to validate our approach, we present results comparing networks synthesized with BNOC with those obtained with an existing benchmarking tool and with already established sets of synthetic networks, in two different scenarios.

Valejo, et al. 2019. A benchmarking tool for the generation of bipartite network models with overlapping communities. *Knowledge and information systems* (accepted, 2019).

Abstract

1) Context: Many real-world networks display hidden community structures with important potential implications in their dynamics. Many algorithms highly relevant to network analysis have been introduced to unveil community structures. Accurate assessment and comparison of alternative solutions are typically approached by benchmarking the target algorithm(s) on a set of diverse networks that exhibit a broad range of controlled features, ensuring the assessment contemplates multiple representative properties.

2) GAP: Tools have been developed to synthesize bipartite networks, but none of the previous solutions address the issue of generating networks with overlapping community structures. This is the motivation for the BNOC tool introduced in this paper.

3) Purpose: It allows synthesizing bipartite networks that mimic a wide range of features from real-world networks, including overlapping community structures. Multiple parameters ensure flexibility in controlling the scale and topological properties of the networks and embedded communities.

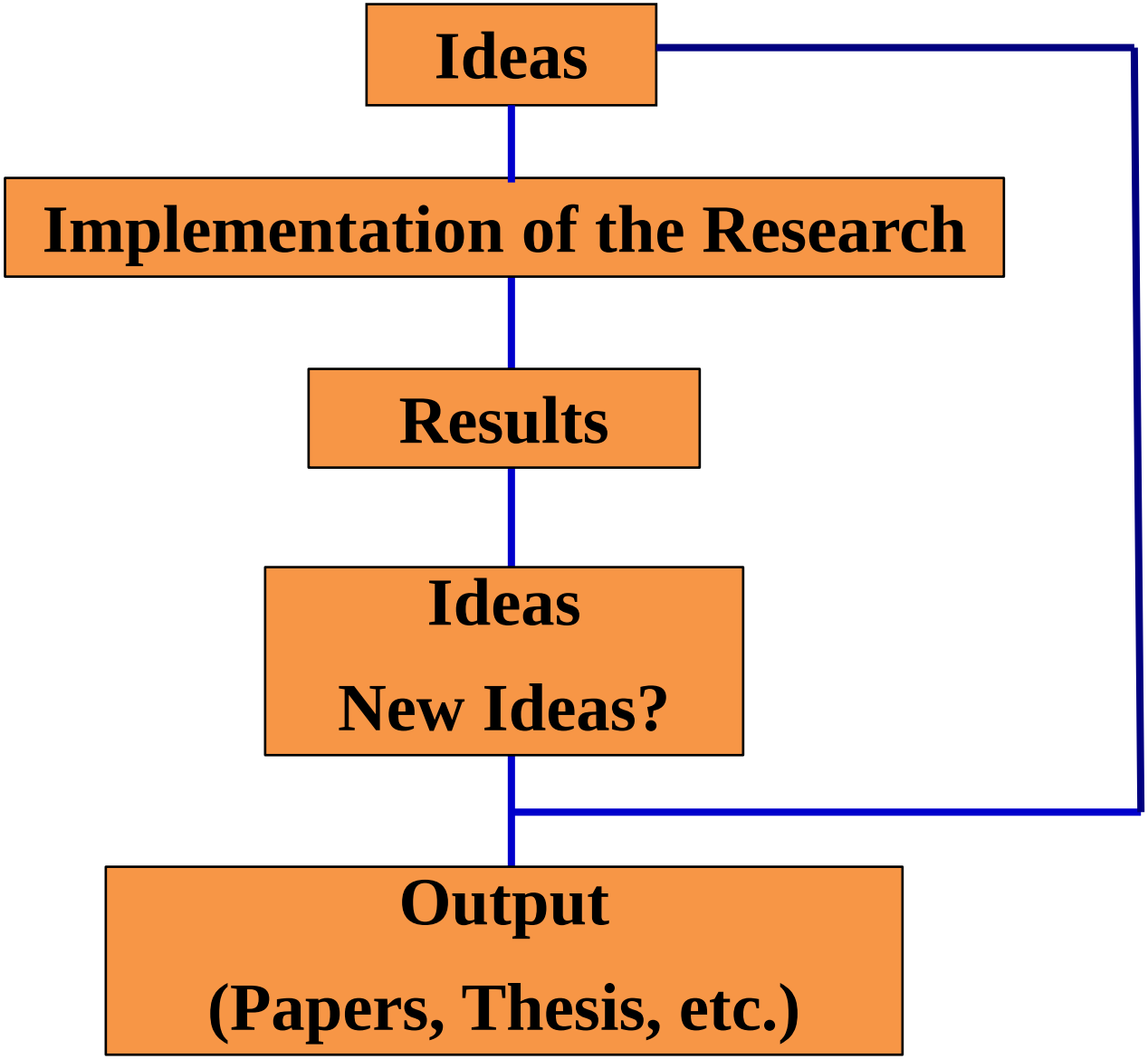
4) Methodology:

5) Results: Results reveal interesting features of the algorithms in this scenario and confirm the relevant role played by a suitable benchmarking tool. Finally, to validate our approach, we present results comparing networks synthesized with BNOC with those obtained with an existing benchmarking tool and with already established sets of synthetic networks, in two different scenarios.

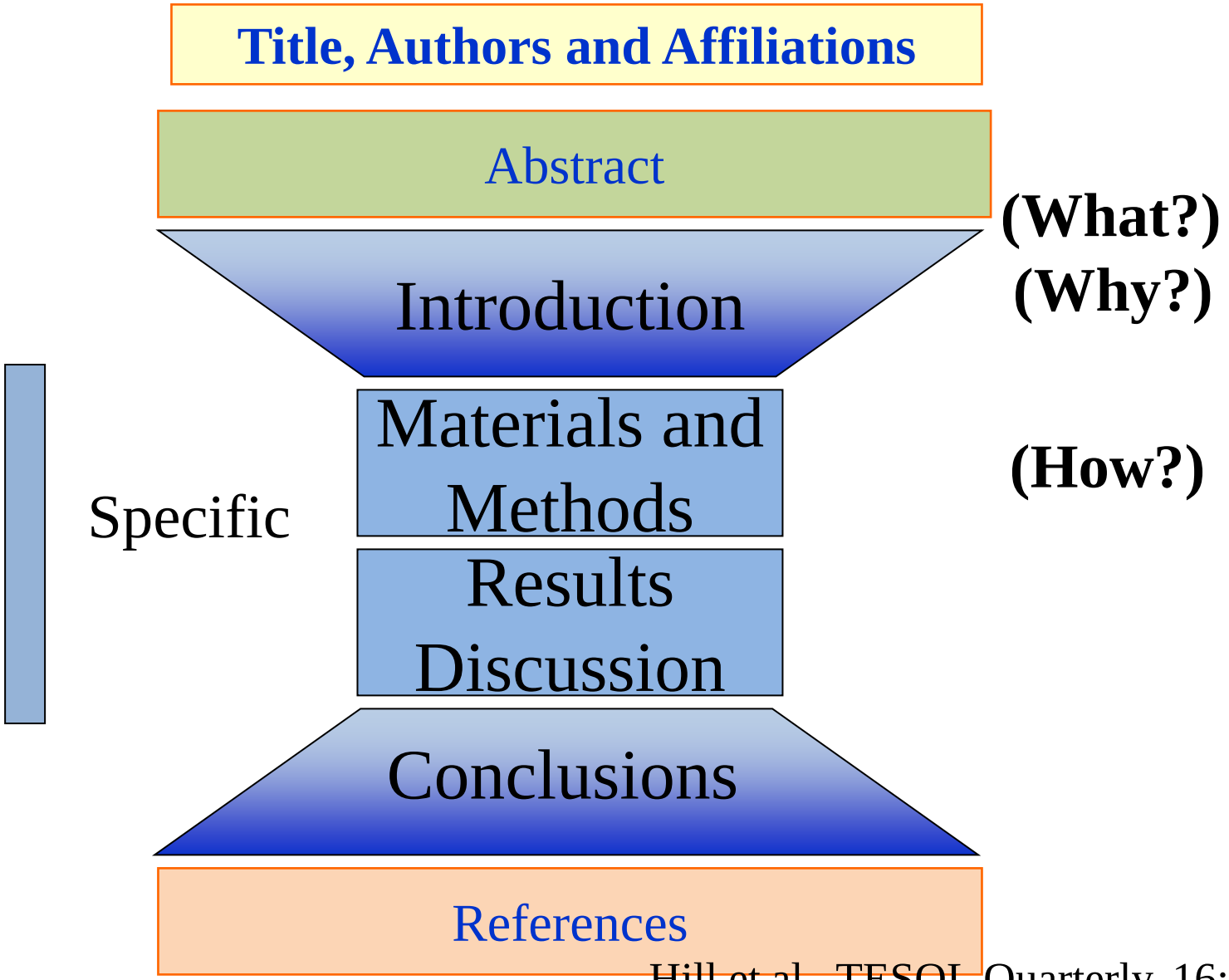
6) Conclusions:

Scientists publish ideas and concepts, NOT results!

Publishing Ideas



Structure of a scientific paper



Hill et al., TESOL Quarterly, 16:333, 1982

Componentes do artigo

- Resumo (fazer no final)
- Introdução (fazer no final)
- Fundamentação teórica, revisão da literatura, tópicos introdutórios
- Metodologia
- Resultados e discussão
- Conclusões

Major difficulties in writing

- Surface errors (typos and grammatical errors)
- Long sentences and short, inadequate paragraphs
- Excessive number of unnecessary words
- Problems in text cohesion, inadequate use of markers and “zig-zag” in the discussion
- Lack of coherent “story” for the text

Major difficulties faced by Brazilian writers (writing in English)

- Lexical
  - Misuse of false cognates and homophone words; lack or misuse of idioms and other collocations employed in scientific texts
- Syntactic
  - Use of grammatical constructions from mother language; word by word translation; over-long/over-complex sentences
- Textual
  - Use of rhetorical structures or strategies of the mother language; misuse of logical relations between sentences or phases; lack of references

First Draft?

- Books and software tools provide help for text post-edition or hints on how to write a paper
  - But normally fail to provide a “hands on” approach that helps the author in producing a first draft.
- To write scientific papers in English it may not be enough:
  - To be fluent in English in another text genre
  - To know the global structure of papers in the mother tongue

A well-written scientific paper and its abstract should follow an underlying organization or structure to convey its content. This means that there are identifiable parts in the paper describing the work. Furthermore, these parts should be organized in an ordered sequence such as:

- Introduction
  1. Problem definition
  2. Previous approaches
  3. Critique: why you/anyone still needs to work on this?
- Contribution (what had as “Gap”): how this addresses 3 above
- Methods (and materials)
- Results
- Conclusions
  1. What has been done and implications
  2. Future work

Tarefa 4 27/09

- Mapear os abstracts dos artigos (dois da Tarefa 3 e outro a sua escolha) na estrutura
  - Context
  - Gap
  - Purpose
  - Methodolgy
  - Results
  - Conclusions
- Informar a ref. completa de cada um

Tarefa 4 27/09

- Para o 2º. artigo escolhido, faça também o mapeamento da estrutura
  - Quais seções correspondem a quais componentes
  - Faça um breve resumo de cada seção

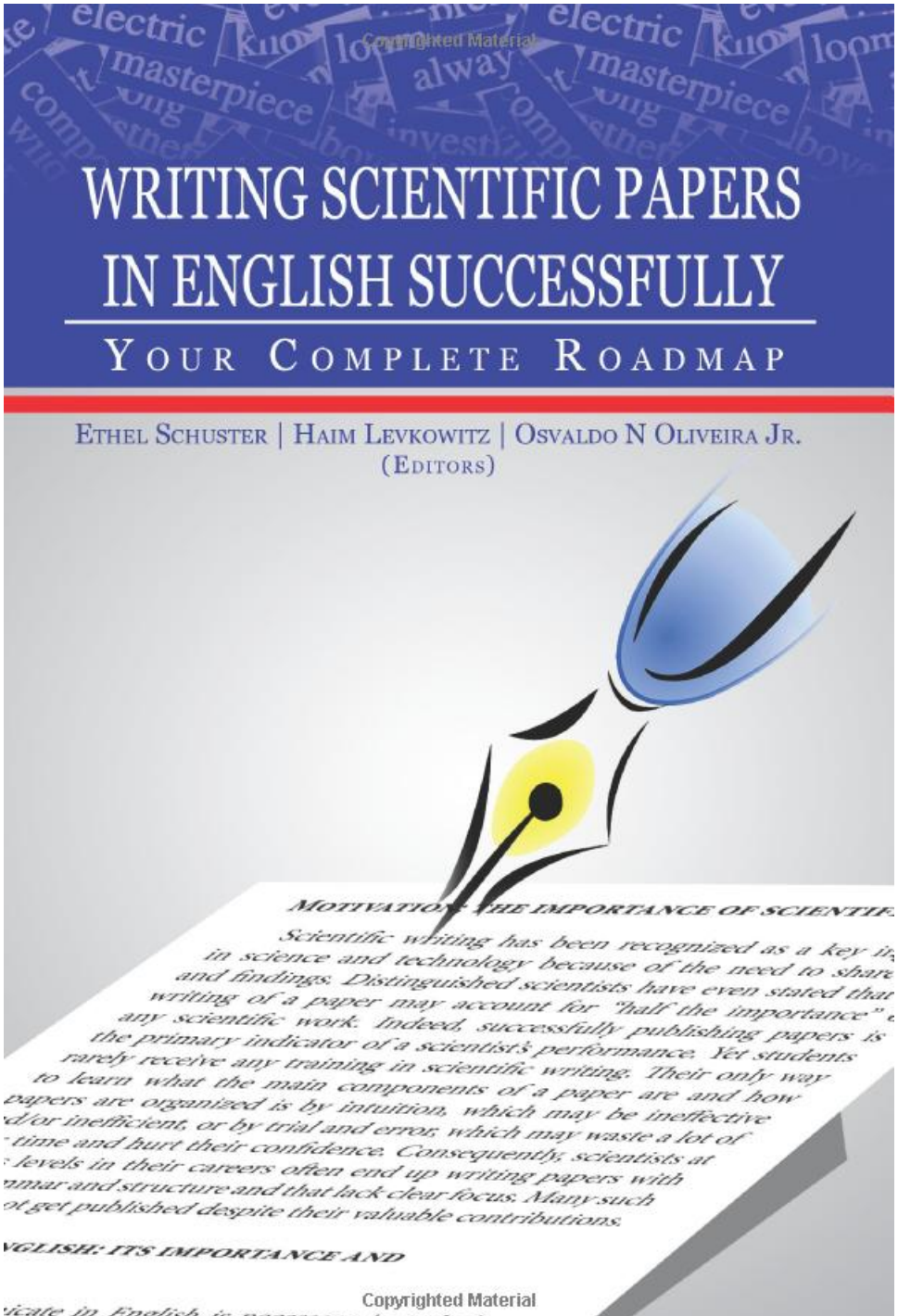
- Enviar por email:
  - Tarefa 4 Metodologia Fulano de Tal

Leituras

- http://www.fapesp.br/boaspraticas/
- http://prp.usp.br/boas-praticas-em-pesquisa/

Fontes e Bibliografia

- Material curso de escrita científica Prof. Osvaldo N. Oliveira Jr. (IFSC)


Copyrighted Material
WRITING SCIENTIFIC PAPERS IN ENGLISH SUCCESSFULLY
YOUR COMPLETE ROADMAP
ETHEL SCHUSTER | HAIM LEVKOWITZ | OSVALDO N OLIVEIRA JR.
(EDITORS)
Scientific writing has been recognized as a key in
in science and technology, because of the need to share
and findings. Distinguished scientists have even stated that
writing of a paper may account for "half the importance."
any scientific work. Indeed, successfully publishing papers is
the primary indicator of a scientist's performance. Yet students
rarely receive any training in scientific writing. Their only way
to learn what the main components of a paper are and how
papers are organized is by intuition, which may be ineffective
d/or inefficient, or by trial and error, which may waste a lot of
time and hurt their confidence. Consequently, scientists at
levels in their careers often end up writing papers with
mmar and structure and that lack clear focus. Many such
ot get published despite their valuable contributions.
Copyrighted Material