Automatic Disambiguation of Author Names: Foundations, Methods and Open Issues

Anderson A. Ferreira[1], **Alberto H. F. Laender**[2]

[1]Departamento de Computação – Universidade Federal de Ouro Preto
Ouro Preto – MG – Brasil

[2]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

anderson.ferreira@ufop.edu.br, laender@dcc.ufmg.br

Abstract. *This tutorial is based on our book "Automatic Disambiguation of Author Names in Bibliographic Repositories" and aims to spread the problem and its challenges among the SBBD community. Author name ambiguity problem occurs when an author publishes works under distinct names or distinct authors publish works under similar names. This problem may be caused by a number of reasons, including the lack of standards and common practices, and the decentralized generation of bibliographic content. In this tutorial, we intend to present an ample view on the automatic disambiguation of author names. We start by discussing its motivational issues, defining the author name disambiguation task and presenting its foundations. Next, we describe some methods proposed by our research group, as well as some recent approaches to author name disambiguation. Finally, we discuss open issues.*

1. Tutorial Identification

1.1. Title

Automatic Disambiguation of Author Names: Foundations, Methods and Open Issues

1.2. Authors

- Anderson Almeida Ferreira (presenter)
  Departamento de Computação – Universidade Federal de Ouro Preto
  anderson.ferreira@ufop.edu.br
- Alberto Henrique Frade Laender
  Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
  laender@dcc.ufmg.br

1.3. Type of Tutorial

Introductory

1.4. Target Audience

The tutorial is primarily target to undergraduate students.

1.5. Presentation Language

Portuguese

1.6. Audiovisual Resources

Notebook and multimedia projector

2. Tutorial Overview

2.1. Previous Presentation

A preliminary version of this tutorial was presented at the 2021 Brazilian Database Symposium (SBBD 2021), in a non-presential way. For this new presentation, which will be held in person, we intend to add some recent works published after that presentation.

2.2. Summary

This tutorial is based on our book "Automatic Disambiguation of Author Names in Bibliographic Repositories" and aims to spread the problem and its challenges among the SBBD community. Author name ambiguity problem occurs when an author publishes works under distinct names or distinct authors publish works under similar names. This problem may be caused by a number of reasons, including the lack of standards and common practices, and the decentralized generation of bibliographic content. In this tutorial, we intend to present an ample view on the automatic disambiguation of author names. We start by discussing its motivational issues, defining the author name disambiguation task and presenting its foundations. Next, we describe some methods proposed by our research group, as well as some recent approaches to author name disambiguation. Finally, we discuss open issues.

2.3. Tutorial Outline

1. Introduction
   (a) Problem Characterization
   (b) Motivational Issues

   In the Introduction, we aim to characterize the author name ambiguity problem and discuss its motivational issues.
2. Foundations
   (a) Basic Definitions
   (b) The Author Name Disambiguation Task
   (c) Evaluation Metrics
3. Taxonomy for Author Name Disambiguation Methods
   In this topic, we start by presenting some basic definitions, followed by a brief description of the main steps involved in the author name disambiguation task. Then, we present some of the evaluation metrics most used in the literature to assess the performance of existing author name disambiguation methods. Finally, we present a taxonomy that we have introduced to classify them according to the type of approach they adopt and the evidence they explore in the disambiguation task.
4. Overview of our Author Name Disambiguation Methods
   (a) HHC: Heuristic-based Hierarchical Clustering
   (b) SAND: Self-training Associative Name Disambiguator
   (c) INDi: Incremental unsupervised Name Disambiguation

In this topic, we illustratively discuss some methods developed by our research group. More specifically, we describe an author name disambiguation method that is based on two specific real-world assumptions regarding scientific authorship, followed by methods based on self-training and an incremental approach.

5. Other Approaches to AND
   We briefly describe some alternative (recent) methods that follow specific approaches that might be considered complementary with respect to those presented in the previous topics.
6. Concluding Remarks and Open Issues
   We conclude the tutorial and discuss some open issues.

3. Brief Professional Biographies of Authors

3.1. Anderson A. Ferreira

Anderson A. Ferreira holds a B.S. degree in Computer Science from the Universidade Federal de Viçosa, Brazil, and an M.Sc. and a Ph.D. degree in Computer Science from the Universidade Federal de Minas Gerais, Brazil, under the supervision of Dr. Marcos André Gonçalves and Prof. Alberto H. F. Laender. In 2011, he joined the Computing Department of the Universidade Federal de Ouro Preto, where he is currently an Associate Professor. He has published several articles in major conferences and journals from the digital libraries and databases areas, such as JCDL, SBBD/JIDM, JASIST, IP&M, DocEng, LA-Web, TKDD, Information Sciences, World Digital, International Journal on Digital Libraries, and SIGMOD Record. Dr. Ferreira has also served as an ad hoc referee for several journals as JASIST, Scientometrics, Information Science, The Knowledge Engineering Review, Informetrics, Online Information Review, IP&M, Internet Services and Applications, Machine Learning Research, KNOSYS, and Natural Language Engineering.

3.2. Alberto H. F. Laender

Alberto H. F. Laender holds a B.Sc. degree in Electrical Engineering (1974) and an M.Sc. degree in Computer Science (1979), both from the Universidade Federal de Minas Gerais, Brazil, and a Ph.D. degree in Computing (1984) from the University of East Anglia, UK. He joined the Computer Science Department of the Universidade Federal de Minas Gerais in 1975, where he is a Full Professor and heads the Data Management Research Group. In 1997, he was a Visiting Scholar at HP Labs in Palo Alto, California. He has served on the advisory committee of several Brazilian research funding agencies and was also a member of ACM SIGMOD's Advisory Board (2006-2010) and SIGMOD's Jim Gray Ph.D. Dissertation Award Committee (2008-2011). Prof. Laender has also served as a program committee member for several national and international conferences on databases, digital libraries, and Web-related topics, among them SBBD, KdMiLe, VLDB, CIKM, SIGIR, JCDL, TPDL, WWW, SPIRE and ICDE. He is a founder-member of the Brazilian Computer Society and one of the co-founders of Akwan Information Technologies, a Brazilian search technology company that was acquired by Google Inc. in 2005 to become its Research and Development Center for Latin America. Prof. Laender is a member of the Brazilian Academy of Sciences and of the Brazilian National Academy of Engineering. In 2010 he was awarded the National Order of the Scientific Merit by the Brazilian President, and in 2022 he received the Scientific Merit Award from the Brazilian

Computing Society and the ER Fellow Award in recognition of his contributions to the International Conference on Conceptual Modeling community. He is the author of more than 200 refereed journal and conference papers. His current research interests include Data Management, Digital Libraries, Social Networks, and Bibliometrics.

Acknowledgments

We thank the support provided by our institutions, UFOP and UFMG, and the research grants received from CAPES, CNPq and FAPEMIG.

References

Boukhers, Z. and Asundi, N. B. (2022). Whois? deep author name disambiguation using bibliographic data. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *Linking Theory and Practice of Digital Libraries*, pages 201–215, Cham. Springer International Publishing.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., and Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, 61(9):1853–1870.

Espiridião, L. V. B., Dias, L. L., and Ferreira, A. A. (2021). Applying data augmentation for disambiguating author names. In *2021: Proceedings of the 36th Brazilian Symposium on Databases, SBBD 2021, Rio de Janeiro, Brazil (Online), October 4-8, 2021*, pages 109–120. SBC.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, 41(2):15–26.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2020). *Automatic Disambiguation of Author Names in Bibliographic Repositories*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., and Laender, A. H. F. (2014). Self-training author name disambiguation for information scarce scenarios. *Journal of the Association for Information Science and Technology*, 65(6):1257–1278.

Hussain, I. and Asghar, S. (2018). Disc: Disambiguating homonyms using graph structural clustering. *Journal of Information Science*, 44(6):830–847.

Liu, Y., Li, W., Huang, Z., and Fang, Q. (2015). A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, 66(3):634–644.

Santana, A. F., Gonçalves, M. A., Laender, A. H., and Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *Journal of the Association for Information Science and Technology*, 68(4):931–945.

Shen, Q., Wu, T., Yang, H., Wu, Y., Qu, H., and Cui, W. (2016). Nameclarifier: A visual analytics system for author name disambiguation. *IEEE transactions on visualization and computer graphics*, 23(1):141–150.