Automatic Methods for Disambiguating Author Names in Bibliographic Data Repositories

Anderson A. Ferreira[1] Marcos André Gonçalves[2] Alberto H. F. Laender[2]
[1]Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto, Brazil
ferreira@iceb.ufop.br
[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{mgoncalv,laender}@dcc.ufmg.br

ABSTRACT

Name ambiguity in the context of bibliographic citation records is a hard problem that affects the quality of services and content in digital libraries and similar systems. This problem occurs when an author publishes works under distinct names or distinct authors publish works under similar names. The challenges of dealing with author name ambiguity have led to a myriad of name disambiguation methods. In this tutorial, we characterize such methods by means of a proposed taxonomy, present an overview of some of the most representative ones and discuss open challenges.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Algorithms, Measurement

Keywords

Ambiguity author names; disambiguation methods

1. INTRODUCTION

Name ambiguity in the context of bibliographic citation records is a hard problem that affects the quality of services and content in digital libraries (DLs) and similar systems. This problem occurs when an author publishes works under distinct names or distinct authors publish works under similar names. The author name ambiguity challenges have led to a myriad of disambiguation methods.

This tutorial is divided in two parts. The first part is based on our survey, entitled "A Brief Survey of Automatic Methods for Author Name Disambiguation", which was published in SIGMOD Record in June 2012 [7]. In this introductory part, we contextualize the problem, present a formal definition for it and propose a general taxonomy for characterizing the automatic author name disambiguation methods proposed in the literature. Then, we briefly describe some of these methods according to our taxonomy and discuss open research challenges. In the second part, we address some of our solutions for the problem. First, we present HHC - Heuristic-based Hierarchical Clustering [3]. HHC disambiguates a set of citation records by successively fusing clusters of citation records with similar author names based on a real-world heuristic applied to their citation attributes. Then, we present SAND - Self-training Associative Name Disambiguator [9, 8]. SAND is a three-step self-training method for author name disambiguation that requires no manual labeling and no parameterization (in real world scenarios). Finally, we present INDi - Incremental unsupervised Name Disambiguation [2]. INDi is an unsupervised incremental method that aims to disambiguate only the new ambiguous citation records inserted into a disambiguated DL. We conclude this second part by presenting SyGAR [6], a tool for generating synthetic collections that allows the simulation of several realistic scenarios to support the evaluation of disambiguation methods.

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.
ACM 978-1-4503-3594-2/15/06.
http://dx.doi.org/10.1145/2756406.2756930 .

2. PROPOSED TAXONOMY

In [7], we proposed a hierarchical taxonomy for grouping the most representative automatic author name disambiguation methods found in the literature. According to our taxonomy, the methods may be classified following the main type of exploited approach: *author grouping* [3, 9, 11, 12, 14, 13, 16, 17], which tries to group the references to the same author using some type of similarity among reference attributes, or *author assignment* [1, 4, 8, 10, 15, 18], which aims at directly assigning the references to their respective authors. Alternatively, the methods may be grouped according to the evidence explored in the disambiguation task: the citation attributes (only), web information, or implicit data that can be extracted from the available information.

3. OVERVIEW OF THE METHODS

In this tutorial, we present an overview of our three most representative methods [5], namely, HHC - Heuristic-based Hierarchical Clustering, SAND - Self-training Associative Name Disambiguator and INDi - Incremental unsupervised Name Disambiguation.

HHC [3] is based on a general heuristic that considers two real-world assumptions: (1) very rarely two authors with similar names and sharing a common co-author are two different persons in the real world and (2) authors tend to publish in the same subjects and publication venues for a considerable portion of their careers. HHC works in two

steps. In the first step, it groups citation records with similar author names into clusters based on the coauthorship relations existing among the records. Then, in the second step, it fuses the initially created clusters based on the similarity of their work or publication venue titles. This process continues until no more clusters can be fused. The final result is a list of clusters with their respective citation records.

SAND [9, 8] is a self-training method that divides the disambiguation task in three steps. Step 1 (*author grouping*) aims to automatically create pure clusters of citation records. SAND obtains pure clusters by exploiting coauthorship relations among the records. SAND considers that a citation record and a cluster share co-authors if both have at least one similar co-author name whose last name is not popular (i.e., it is not a common last name) or if they have at least two similar co-author names (popular or not). Step 2 (*cluster selection*) aims to select, from clusters produced by Step 1, the ones belonging to distinct authors for composing the training data used by Step 3. We select the dissimilar clusters with the largest number of citation records. The citation records from these selected clusters are inserted into the training data D, along with the author label of the corresponding cluster. Step 3 (*author assignment*) uses the set of examples D to produce a disambiguation function for predicting the correct author of the unselected clusters, based on a lazy associative classifier [19]. SAND detects authors without a representative cluster in the training data D and includes them in D. It also increases the coverage of D by exploiting reliable predictions.

INDi [2] is an incremental author name disambiguation method aimed at determining the authors of new citation records as they get added to a DL. As such, it need not be applied to the whole DL at once to disambiguate it, thus preserving eventual manual corrections previously done. INDi includes specific heuristics to check whether author names of new citation records refer to pre-existing authors in the DL or to new ones (authors with no citation records in the DL). These heuristics are meant to disambiguate new citation records by prioritizing the assignment of such records to the correct author instead of assigning the doubtful record to an existing author with a probability of error. INDi attempts to disambiguate the new citation records by looking for an existing author whose records in the DL include a similar author name, at least one common co-author, and similar work or publication venue titles. For cases in which the new citation record does not include co-authors or all existing records in a group of an existing similar author do not include any co-authors, it does not perform the co-author check, but raises similarity thresholds for publication venue and work title. When this checking procedure fails, the citation record is considered as belonging to a new author.

4. CONCLUDING REMARKS

Author name ambiguity is a hard problem that affects the quality of services and content in DLs. Thus, at the end of this tutorial, we expect the participants will have understood the challenges of this problem and learned about some of the existing solutions.

Acknowledgments

This work has been supported by InWeb and by the authors' individual grants from CAPES, CNPq and FAPEMIG.

References

[1] I. Bhattacharya and L. Getoor. A Latent Dirichlet Model for Unsupervised Entity Resolution. In *SDM*, 2006.

[2] A. P. Carvalho, A. A. Ferreira, A. H. F. Laender, and M. A. Gonçalves. Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries. *JIDM*, 2(3):289–304, 2011.

[3] R. G. Cota, A. A. Ferreira, M. A. Gonçalves, A. H. F. Laender, and C. Nascimento. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *JASIST*, 61(9):1853–1870, 2010.

[4] L. V. B. Esperidião, A. A. Ferreira, A. H. F. Laender, M. A. Gonçalves, D. M. Gomes, A. I. Tavares, and G. T. de Assis. Reducing Fragmentation in Incremental Author Name Disambiguation. *JIDM*, 5(3):293–307, 2014.

[5] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender. Disambiguating author names using minimum bibliographic information. *World Digital Libraries*, 7(1):71–84, 2014.

[6] A. A. Ferreira, M. A. Gonçalves, J. M. Almeida, A. H. F. Laender, and A. Veloso. A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Information Sciences*, 206:42–62, 2012.

[7] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender. A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record*, 41(2):15–26, 2012.

[8] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender. Self-training author name disambiguation for information scarce scenarios. *JASIST*, 65(6):1257–1278, 2014.

[9] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender. Effective self-training author name disambiguation in scholarly digital libraries. In *JCDL*, pages 39–48, 2010.

[10] H. Han, C. L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. In *JCDL*, pages 296–305, 2004.

[11] J. Huang, S. Ertekin, and C. L. Giles. Efficient Name Disambiguation for Large-Scale Databases. In *PKDD*, pages 536–544, 2006.

[12] I.-S. Kang, S.-H. Na, S. Lee, H. Jung, P. Kim, W.-K. Sung, and J.-H. Lee. On co-authorship for author disambiguation. *IPM*, 45(1):84–97, 2009.

[13] A. F. Santana, M. A. Gonçalves, A. H. F. Laender, and A. A. Ferreira. Combining Domain-specific Heuristics for Author Name Disambiguation. In *JCDL*, pages 173–182, 2014.

[14] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles. Efficient Topic-based Unsupervised Name Disambiguation. In *JCDL*, pages 342–351, 2007.

[15] J. Tang, A. C. M. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *TKDE*, 24(6):975–987, 2012.

[16] V. I. Torvik and N. R. Smalheiser. Author name disambiguation in medline. *TKDD*, 3(3):1–29, 2009.

[17] P. Treeratpituk and C. L. Giles. Disambiguating Authors in Academic Publications using Random Forests. In *JCDL*, pages 39–48, 2009.

[18] A. Veloso, A. A. Ferreira, M. A. Gonçalves, A. H. Laender, and W. Meira Jr. Cost-effective on-demand associative author name disambiguation. *IPM*, 48(4):680 – 697, 2012.

[19] A. Veloso, W. M. Jr., M. A. Gonçalves, and M. J. Zaki. Multi-label lazy associative classification. In *PKDD*, pages 605–612, 2007.