Using Co-authorship Networks for Author Name Disambiguation

Fakhri Momeni
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
fakhri.momeni@gesis.org

Philipp Mayr
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
philipp.mayr@gesis.org

ABSTRACT

With the increasing size of digital libraries (DLs) it has become a challenge to identify author names correctly and assign publications to them. The situation becomes more critical when different persons share the same name (homonym problem) or when the names of authors are presented in several different ways (synonym problem). This paper focuses on homonym names in the computer science bibliography DBLP. The goal of this study is to implement and evaluate a method which uses co-authorship networks in order to disambiguate homonym names, especially common names. The results show that the implemented method has a good performance and can be used for author name disambiguation of sparse bibliographic records.

Keywords

Author name homonyms; Co-authorship network; Community detection; Louvain method; Gold standard

1 Introduction

In scholarly digital libraries authors are recognized via their publications. It is important for users to know about the author of a particular publication to access possible other publications by this author. For this purpose DLs provide search services by using the publication information in their databases. However, when several authors share the same name or authors provide their works under different versions of their name, DLs need more analysis on authors' oeuvres. Manual author identification in large DLs is very costly. Thus, as a consequence, automated solutions are to be found to analyze large sets of ambiguous author names. In addition, the demographic characteristics such as name origin and frequency of names used for authors influence the identification of authors. Therefore, all constraints of the underlying data should be considered to choose the appropriate method for author name disambiguation.

The author assignment and author grouping methods [3] are the two main types of method for author name disambiguation. The author assignment methods construct a model that represents the author and assigns proper publications to the model. It requires former knowledge about the authors. Nguyen and Cao [6] used these methods and proposed to link the author names to the matching entities in Wikipedia. The author grouping methods cluster the publications on the basis of their properties (co-authors, publication year, keywords, etc.) to assign a group of publications to a certain author. Following this framework, Caron and van Eck [2] applied rule-based scoring to clustered publications. In their approach they suppose that there is enough information about authors and their documents. Also, Gurney et al. [4] clustered publications with employing different data fields and integrated a community detection method.

In this paper we used an author grouping method (compare [3]) to cluster the publications of a set of random authors with the same name in the DBLP database. Considering the lack of rich bibliographic information in DBLP records, we applied co-authorship network analysis to detect similarities between publications. In addition, we investigated how the amount of homonym names affects the disambiguation results. In the end, we employed a community detection algorithm (Louvain method) to reduce the effect of common names in our evaluation.

2 Disambiguation Approach

We use an author grouping method in order to assign all publications of each person to a certain group. For this purpose all publications belonging to the same ambiguous author name are categorized into one block. In a next step we compare every pair of publications in each block with each other to find a similarity between them. If we have n blocks and m_i publications in a block i, the number of comparisons for all blocks is:

$$\sum_{i=1}^{n} \frac{m_i(m_i-1)}{2} \quad (1)$$

The result of each comparison is true or false. The *true* result means that two publications belong to one person and the same cluster. If one of them was compared with another one before and assigned to a cluster, the other one is added to that cluster too. If both of them were compared before and belong to different clusters, two clusters are rebuilt to one cluster. Otherwise a new cluster will be created and two publications are put in new cluster. In the next section we describe how to define the similarity indicator to build the clusters. The bibliographic information that we can obtain from publications in DBLP is limited mainly to author names (the names of all co-author names are listed), title

JCDL '16 June 19-23, 2016, Newark, NJ, USA

ACM ISBN 978-1-4503-4229-2/16/06.

DOI: http://dx.doi.org/10.1145/2910896.2925461

and publication venue. We chose the co-author names as our similarity indicator. Therefore, we built a network of all authors and documents. Each pair of documents within every block has to be compared. To compare the publications the relations in the network are analyzed. If there is a path between two publications, their distance is defined as the length of the shortest path between them, otherwise it would be infinite. The length of the shortest path is equal to the number of nodes between two nodes. The less distance between two publications meansthat these publications were more likely to be written by one person. So, the distance between two publications is assumed as the similarity measure. Different thresholds can be considered for the distance.

3 Evaluation

In order to evaluate the output of the author disambiguation approach we need a gold standard of disambiguated author names. Many homonym author names in DBLP are disambiguated manually by the DBLP team and are identifiable with an id. For example, 'Hui Lin' belongs to four different persons: 'Hui Lin 0001', 'Hui Lin 0002', 'Hui Lin 0003' and 'Hui Lin 0004'. Thus, the set of publications for each person is recognizable. To build the gold standard[1] [5] we selected these identified author names and compiled all their publications into one set. In our gold standard we provide a list of publications that have at least one disambiguated author name. There are 5,408 authors who have an identification number (we mention them as disambiguated authors). These 5,408 authors and their publications form the gold standard.

To measure the performance of our method 1,000 disambiguated author names have been randomly selected from the gold standard. In total we have 2,844 different authors (with calculating their identifier) and 32,273 publications in our random sample. In the next section we evaluate the performance of our method against the gold standard. Some authors report that metrics like precision and recall have some constraints proving that are not suitable for evaluation of the effectiveness of clustering algorithms (see e.g. [1]). *BCubed* precision and recall [1] are metrics that satisfy these constraints and therefore we applied them to evaluate our method. For this purpose BCubed precision and recall were computed for each publication. The publication precision measures how many publications in its group belong to its author. The publication recall measures how many publications from its author appear in its group. BCubed precision, recall and F-measure were computed for every publication in every block. Then we considered their average as the BCubed precision, recall and F of the block.

4 Results and Discussion

For choosing the threshold we have checked the distances larger than 3, which results in a very low precision. Then we chose the threshold equal to 1 and 3. For the distance equal or less than threshold (1 or 3), we assign two publications in the same cluster. The results of the evaluations for two thresholds are demonstrated in Table 1.

Table 1: Mean values of BCubed metrics for 1,000 blocks

	BCubed precision	BCubed recall	BCubed F
Threshold=1	0.99	0.77	0.81
Threshold=3	0.96	0.83	0.84

The results in Table 1 indicate that our co-author networks method performs well on the dataset and it can be utilized as author identification approach. Comparing the results for two thresholds (1 and 3) we can conclude that using threshold = 3 provides us with the better balance between precision and recall and a higher F (slightly better BCubed recall of 0.83 and F of 0.84). We observed that although using threshold=3 results the better performance generally, it is less efficient than using threshold=1 for common names. The reason is that common names enhance the probability of being authors with the same name in the same area of research activity and increase the likelihood of detecting the shared co-author for different researchers with the same name. Furthermore, it is more likely that these authors have co-authors with similar common names. This results in a higher probability of ambiguous co-authors and wrong connections between publications. Therefore, we should be more cautious when using the co-author of co-author as the similarity measure for these cases and verify the results more deeply. Hence, we applied a community detection algorithm to optimize the results (threshold=3) for the common names. We chose a subset of the names which have more than 200 publications (in total 28 names) in our DBLP dataset. To detect communities in the network we utilized the Louvain method with Pajek.

Because this method is based on co-author network, it is limited to multi-author papers. Therefore, a multi-aspect indicator is required for single-author papers. In this way, we can use the titles of publications to extract keywords and use this information to calculate similarity measures.

[1] Available at http://dx.doi.org/10.7802/1234

5 References

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486, 2009.

[2] E. Caron and N. J. van Eck. Large scale author name disambiguation using rule-based scoring and clustering. pages 79–86, 2014.

[3] 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.

[4] T. Gurney, E. Horlings, and P. V. den Besselaar. Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91(2):435–449, 2012.

[5] P. Mayr and F. Momeni. An open testbed for author name disambiguation evaluation.

[6] H. T. Nguyen and T. H. Cao. Named entity disambiguation: A hybrid statistical and rule-based incremental approach. In *The Semantic Web, 3rd Asian Semantic Web Conference, ASWC 2008, Bangkok, Thailand, December 8-11, 2008. Proceedings*, pages 420–433, 2008.