2021 IEEE 37th International Conference on Data Engineering (ICDE) | 978-1-7281-9184-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICDE51399.2021.00244

A Collective Approach to Scholar Name Disambiguation

(Extended Abstract)

Dongsheng Luo*, Shuai Ma†, Yaowei Yan*, Chunmin Hu†, Xiang Zhang*, Jinpeng Huai†
*Pennsylvania State University
†SKLSDE Lab, Beihang University, Beijing, China
†Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China
{dul262, yxy230, xzz89}@psu.edu, {mashuai, hucm, huaijp}@buaa.edu.cn

***Abstract*—This study investigates name disambiguation for scholarly data. We propose a collective approach, which considers the connections of different ambiguous names, such that it initially treats each author reference as a unique author entity and reformulates the bibliography data as a heterogeneous multipartite network. Disambiguation results of one author name propagate to the others in the network. To further deal with the sparsity problem caused by limited available information, we also introduce word-word and venue-venue similarities and measure author similarities by assembling similarities from multiple perspectives. Using three real-life datasets, we experimentally show that our approach is both effective and efficient.**

I. Introduction

Scholar name ambiguity is a common data quality problem for digital libraries and has raised various troubles in scholar search, document retrieval, and so on. Limited information available in bibliography data makes this problem more challenging to attack. Most existing methods tackle name disambiguation separately and independently [1]–[4]. However, neglecting their connections may lead to sub-optimal solutions.

To this end, we propose a collective approach, referred to as NDCC, to dealing with scholar name disambiguation using only the limited information common available for bibliography data. A heterogeneous multipartite network is adopted to represent the dataset. Disambiguation results of one name affect the others by updating the graph structure. To tackle the sparsity data problem, we develop a novel metric for determining the author similarity by assembling the similarities of four features (*i.e.,* coauthors, venues, titles, and coauthor names) available in bibliography data. Comprehensive experimental studies on three real-life datasets show that our method NDCC is both effective and efficient.

The full version of this extended abstract appears in [5].

II. Problem Formulation

We use a heterogeneous multipartite network $\mathcal{G}$ to model a bibliography data D. By considering each author reference as a unique author entity initially, $\mathcal{G}$ contains the sets of author nodes (A), paper nodes (P), venue nodes (V) and title word nodes (T). There are three types of edges in this network, *i.e.,* edges connecting author nodes to paper nodes, paper nodes to venue nodes, and paper nodes to word nodes.

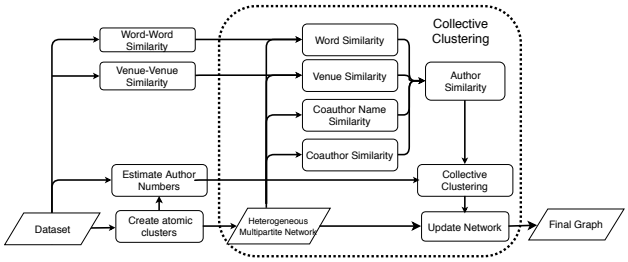

Fig. 1. Framework of NDCC

The corresponding adjacency matrices are $\mathbf{W}^{AP}$, $\mathbf{W}^{PV}$ and $\mathbf{W}^{PT}$, respectively. The task of scholar name disambiguation is to adjust author nodes and edges between author and paper nodes, such that for each author a in A, the set of paper nodes P_a connected to a ideally contains all and only those papers written by author a.

III. Solution Framework

The solution framework NDCC is illustrated in Fig. 1. We differentiate coauthors from coauthor names and determine the author similarity by assembling the similarities from four perspectives (coauthors, venues, titles, and coauthor names). To alleviate the sparsity problem, words used by authors are extended by considering the words similar to their title words, so do venues. The venue-venue and word-word similarities are computed as a preprocessing step. Considering the mutual influence between name disambiguation processes for different names, we propose a bottom-up collective clustering method. The disambiguation of one name affects others by changing the structure of the heterogeneous multipartite network. We iteratively select an author name and calculate the pairwise similarities of its author nodes. We then merge the pairs of author nodes with high similarity scores and update the network accordingly. Each name needs to be disambiguated several times until it is fully disambiguated. To determine the stop condition, we estimate the number of authors for each name. A name is considered to be fully disambiguated if the number of its author nodes reaches the estimated number.

978-1-7281-9184-3/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDE51399.2021.00244

IV. Author Similarity Measurement

We introduce the preprocessing step to deal with the sparsity problem. Then a novel metric is proposed to assemble the similarities from multiple perspectives.

Dealing with Sparsity. Some authors only connect to a small number of paper nodes, especially in the initial heterogeneous multipartite network. It is hard to make a good judgment for these authors. To deal with this sparsity problem, we introduce word-word and venue-venue similarities to expand the limited information. Word-word similarity scores are measured by their word embeddings. Venue-venue similarity scores are calculated by the Jaccard index of their authors

Author Similarity Assembling. The author similarity is assembled by four similarities (coauthors, venues, titles, coauthor names). Given two authors i and j with the same name n, we consider each pair of perspectives and define the author similarity as $sim = \sqrt{\sum_{x \neq y} sim_x \times sim_y}$, where $x, y \in \{n, t, v, a\}$ and $sim_a, sim_n, sim_t, sim_v$ are coauthor, coauthor name, title and venue similarities, respectively. The normalized histogram intersection kernel is utilized to calculate the coauthor similarity sim_a:

$$sim_a = \sum_k \frac{1}{\mathbf{d}_k^A} \min(\mathbf{W}_{i,k}^{AA}, \mathbf{W}_{j,k}^{AA}) + t(n)\{\sum_k \frac{1}{\mathbf{d}_k^A} \min(\mathbf{W}_{i,k}^{AA}, \mathbf{W}_{j,k}^{AA^2}) + \sum_k \frac{1}{\mathbf{d}_k^A} \min(\mathbf{W}_{i,k}^{AA^2}, \mathbf{W}_{j,k}^{AA}))\}. \quad (1)$$

Here matrices $\mathbf{W}^{AA}$ and $\mathbf{W}^{AA^2}$ store valid coauthorship and 2-hop coauthorship, respectively, $\mathbf{d}_k^A$ is the number of papers written by author k, served as the normalization factor. $t(n)$ determines whether to use multi-hop coauthorships. The other similarity scores, *i.e.,* coauthor name, title, and venue similarities, are computed in similar ways.

V. Collective Clustering

We first adopt a highly restrictive rule that two author references are assigned to the same atomic clusters if they share at least two coauthor names, to generate atomic clusters. It significantly reduces the size of the initial network. A statistical method is utilized to estimate the number of authors for each name, which serves as the stop condition.

In collective clustering, disambiguation of one name affects the others by updating the structure of the heterogeneous multipartite network $\mathcal{G}$. In each iteration, we select a name n, calculate the pairwise author similarities with that name, and merge the top K pairs with the highest similarity scores. Here we choose K as the half of the difference between the current author number and the estimated one. Each name is disambiguated iteratively until it is fully disambiguated, *i.e.,* the number of its authors reaches the estimated number. The final network is the disambiguation result.

For efficiency, we calculate and store matrices such as $\mathbf{W}^{AA}$ and $\mathbf{W}^{AA^2}$ as a preprocessing step before iterations, and update them inside iterations. Considering the sparsity and dynamics of matrices, we use lists of treemaps to store these matrices. For each author name, we maintain a list of its author nodes. Each author node contains six treemaps to store the corresponding rows in these metrics, respectively.

Theoretically, the iteration number is no more than $|N|(\log(\ell) + 2)$, where N is the set of author names and ℓ is the largest number of papers written by the authors with the same name. The time complexity is $O(\ell^2 \log(\ell)(\sum_n |A_n^{(0)}|^2 + |A^{(0)}| \log(\ell)))$, where $A_n^{(0)}$ is set of atomic authors with name n, and $A^{(0)} = \bigcup A_n^{(0)}$. The space complexity is $O(|A^{(0)}|\ell^2)$.

VI. Experimental Study

We present an extensive experimental study using three real datasets AMiner, ACM and DBLP. The test set (https://aminer.org/ disambiguation) contains $6,730$ labeled papers of 110 author names. Five methods CE [6], GHOST [1], CSLR [3], MIX [2] and AM [4] are compared.

Experimental Results. (1) Our approach NDCC is effective for scholar name disambiguation. NDCC on average improves Macro-F1 over (CE, GHOST, CSLR, MIX, AM) by (17.87%, 23.25%, 16.65%, 45.39%, 21.24%) on AMiner, (25.36%, 24.26%, 14.16%, 37.46%, 14.96%) on ACM, and (13.11%, 23.31%, 8.47%, 50.37%, 9.86%) on DBLP, respectively.

(2) NDCC is very efficient. With speeding up strategies, NDCC is on average (18, 195, 19) times faster than (CE, CSLR and MIX) on AMiner, (15, 8) times faster than (CE, MIX) on ACM, and 10 times faster than MIX on DBLP, respectively.

(3) Strategies dealing with sparsity improve the accuracy. Incorporating word-word and venue-venue similarities improves Macro-F1 by (0.59%, 2.21%, 0.26%) and (7.28%, 4.58%, 3.09%) on (AMiner, ACM, DBLP), respectively.

VII. Conclusions

Considering the connections of scholar names, we have proposed a collective approach to scholar name disambiguation. We have developed a novel metric to determine the author similarity by assembling the similarities of multiple features. To deal with the sparsity problem, we have also introduced word-word and venue-venue similarities. Extensive experiments have shown that NDCC is both effective and efficient for scholar name disambiguation.

Acknowledgments

This work is supported in part by National Key R&D Program of China 2018AAA0102301 and NSFC 61925203 & U1636210. For any correspondence, please refer to Shuai Ma.

References

[1] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On graph-based name disambiguation," *JDIQ*, vol. 2, no. 2, pp. 10:1–10:23, 2011.

[2] M. Khabsa, P. Treeratpituk, and C. L. Giles, "Online person name disambiguation with constraints," in *JCDL*, 2015.

[3] S. Li, G. Cong, and C. Miao, "Author name disambiguation using a new categorical distribution similarity," in *ECML/PKDD*, 2012.

[4] Y. Zhang, F. Zhang, P. Yao, and J. Tang, "Name disambiguation in aminer: Clustering, maintenance, and human in the loop." in *SIGKDD*, 2018.

[5] D. Luo, S. Ma, Y. Yan, C. Hu, X. Zhang, and J. Huai, "A collective approach to scholar name disambiguation," *TKDE*, online, 2020.

[6] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *TKDD*, vol. 1, no. 1, p. 5, 2007.