Detecting Ambiguous Author Names in Crowdsourced Scholarly Data

Xiaoling Sun*†, Jasleen Kaur†, Lino Possamai‡ and Filippo Menczer†
*Department of Computer Science and Technology
Dalian University of Technology, Dalian, China
†School of Informatics and Computing
Indiana University, Bloomington, USA
‡Department of Pure and Applied Mathematics
University of Padua, Italy
{sun20, jakaur, fil}@indiana.edu, lino@possamai.it

***Abstract*—The name ambiguity problem is a challenge in many areas, especially in the field of bibliographic digital libraries. For example, in services that use citation data to compute the impact of authors, ambiguous names lead to biased measures. The problem is amplified where names are collected from heterogeneous sources, including crowdsourced annotations. This is the case in the *Scholarometer* system, which cross-correlates author names in user queries with those retrieved from bibliographic data. The uncontrolled nature of user-generated annotations is very valuable, but creates the need to detect ambiguous names. In this paper, we propose an approach to detect ambiguous names at query time, which makes it applicable in the context of a social computing application. We explore two kinds of heuristic features based on citations and crowdsourced topics. Our approach can detect ambiguous author names in crowdsourced scholarly data with an accuracy of 75%.**

***Keywords*-Ambiguous name detection; crowdsourcing; citation analysis; scholarly data; discipline annotations; social tagging**

I. Introduction

Bibliometric methods measure the impact of researchers, journals, papers, etc. In a quest for quantitative impact analysis of an author, a wealth of measures based on citation data have been proposed. These measures rely on citations; consistent, accurate and up-to-date citation data is therefore critical for an accurate assessment of author impact.

Usually the publications of an author are identified by the author's name. In reality, however, we cannot always correctly map publications to authors because names can be ambiguous. There are basically two types of name ambiguities: an author may have multiple name variations or multiple authors may share the same name [1]. Name ambiguity can affect the accuracy of citation-based impact analysis and methods to detect ambiguous names are needed.

Scholarometer (scholarometer.indiana.edu) is a social tool to facilitate citation analysis and help evaluate the impact of authors [2]. In our social approach to scholarly citation analysis, information that is crowdsourced from end-users of the system forms the very basis for the service provided. Some of this information comes directly from the users, who provide discipline annotation (tags) for queried authors. Other information is obtained indirectly as a side effect of user queries, by cross-correlating the name of the queried author with bibliographic and citation data retrieved from a digital library. Google Scholar is used as the source for Scholarometer's bibliographic and citation data. However, the name ambiguity problem is independent of the bibliographic data sources.

Prior work in name disambiguation [1], [3]–[11] is based on expensive supervised or unsupervised machine learning algorithms that partition a set of publications into coherent subsets. However, none of these approaches is applicable in the context of social citation analysis tools, which require query-time, real-time detection of ambiguous names. If a name is deemed ambiguous, the tool can ask the user to refine the query, for example by adding keywords to make the query more specific. As a result, we can obtain cleaner data from the user as well as reliable data from the digital library, thus improving the assessment of scholarly impact. Our definition of ambiguity is therefore based on user queries, giving rise to the following formulation of our problem: *Given a set of publications, we need to decide if the author names of these publications match the name in the query*.

Contributions and Outline

Our aim in this paper is to detect ambiguous names in citation impact analysis at query time, by using different kinds of features. The remainder of the paper is organized as follows. After background on related work in § II, we describe the proposed approach in § III, which includes the following contributions:

- A heuristic based on name variations and citations, useful when there are multiple name variations in an author's publications (§ III-A).
- An algorithm to measure the consistency between the topics associated with publication metadata with the help of crowdsourced discipline annotations (§ III-B).

These methods yield features that can be used to estimate the likelihood that a set of publications belong to the same author. In § IV we evaluate these features in a supervised learning setting and show the effectiveness of combining them.

II. Related Work

The name ambiguity problem has been studied in a wide variety of contexts, such as biomedical term disambiguation [12], geographic name disambiguation [13], and personal name disambiguation [14]. Especially in the context of bibliographic citation records, researchers have proposed numerous methods for author name disambiguation within bibliographic databases and on the Web [1], [3]–[11], [15]. The problem is important because it affects the quality of content and services in digital libraries. In prior work, name disambiguation has been cast into the problem of clustering a set of publications into profiles such that each profile corresponds to a single author.

The literature on disambiguation is mainly categorized into supervised and unsupervised learning approaches. Supervised learning approaches [5], [7], [15] use a set of authors with given partitions to train a classifier to recognize whether two publications, or two sets of publications, belong to the same profile. Unsupervised approaches [1], [3], [4], [6], [8]–[11] do not use training examples but instead exploit publication features to merge similar publications into clusters. A variety of clustering algorithms have been explored and compared for unsupervised name disambiguation.

The approach taken by the Scholarometer system is also based on supervised learning, but defines the disambiguation problem in a different way. Given a set of papers for a given author name, the task is to determine whether the name is ambiguous, i.e. corresponds to multiple authors. This is consistent with the social nature of the tool, which aims to leverage quality data provided by users. Our first attempt to deal with ambiguous authors names deployed a simple heuristic based on name variations and citations [2]. This heuristic will be discussed in § III-A. With the increasing popularity of the tool, the number of authors in the Scholarometer system has grown significantly, revealing that many ambiguous names were undetected. This has motivated the approach presented here.

III. Ambiguous Author Name Detection

The goal of the proposed approach for author name disambiguation is to detect an ambiguous author name when it is submitted as a query to the citation analysis system. To the best of our knowledge, this is the first attempt to cast name disambiguation this way. Our algorithm extracts features from all publications retrieved for the queried author name, and performs binary classification. In this section we describe two classes of features of the publications of a scholar that will be supplied to our classifier.

A. Name Variations and Citations

We extract the name variations from a collection of publications returned by a bibliographic digital library (Google Scholar in our case) for the query and sort them by number of citations. Typical author names have two or three variations (e.g., with and without a middle initial). Therefore we look at the percentage of the total citations that are attributed to the top name variations. A high percentage suggests that the name is not ambiguous, as any further variations only account for a small fraction of the citations and therefore do not have a large effect on impact measures. On the other hand, if the top name variations account for a low fraction of the citations, we assume that the name is ambiguous.

The heuristic approach first used to detect ambiguous names in the Scholarometer system was based on this idea, and used a fixed threshold (90%) for the fraction of citations to papers corresponding to the *top three* name variations [2]. To apply the same idea in the supervised learning setting, we use the fraction of citations for the top n ($n = 1, 2, 3$) name variations as a feature for the classifier. We call such a feature *citation per name variation* (CNV_n). Of course this is a useful feature only when there are a sufficient number of name variations in the set of publications.

B. Topic Consistency

In this subsection we attempt to capture the topical consistency among a set of papers of an author. The algorithm makes use of topics associated with publications and similarities among them. To this end, we leverage the discipline tags crowdsourced from the users of the Scholarometer system. Each discipline can be represented by a vector of authors who have been tagged with it. A vector weight is the number of times that an author has been tagged with a discipline. The similarity σ between two disciplines, calculated by cosine similarity, reflects the strength of cross-disciplinary collaborations between authors in these two disciplines. We list ten pairs of highly related disciplines in Table I.

With the growth of interdisciplinary research collaborations, the research topics of an author are not limited to one discipline. More and more authors are publishing papers in different disciplines. To capture the possibility that publications in different disciplines may be from the same author in spite of inconsistent metadata, we need to detect different but related disciplines associated with an author name. For example, an author has two subsets of publications, A and B, in different disciplines. Suppose that the publication metadata is consistent within each subset, but not between A and B. If we know that publications in A are about topic 1, publications in B are about topic 2, and topic 1 and topic 2 are highly related, we may infer that the publications are consistent and the author is not ambiguous. This is the basic intuition for a feature that we call *publication topic consistency* (PTC).

Based on the above intuition, we need to map an author's publications to topics, and measure the similarity between these topics. We assume that any publications can be mapped to one or more topics from a preexisting set. In the case of the Scholarometer system, we use a subset of the crowdsourced tags from a controlled vocabulary, namely the Journal Citation Reports (JCR) categories. These 242 categories are composed of Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index from the Web of Science, an academic citation index provided by Thomson Reuters.[1] For every author, all the publication titles and venues

[1] en.wikipedia.org/wiki/Web_of_Science

TABLE I
TEN HIGHLY RELATED DISCIPLINE PAIRS AND THEIR COSINE SIMILARITIES

computer science, hardware & architecture	engineering, electrical & electronic	0.95
materials science, multidisciplinary	optics	0.82
neurosciences	language & linguistics	0.78
surgery	transplantation	0.75
geology	paleontology	0.73
behavioral sciences	biodiversity conservation	0.70
demography	management	0.70
energy & fuels	operations research & management science	0.65
anatomy & morphology	obstetrics & gynecology	0.63
ecology	paleontology	0.63

TABLE II
TOP 10 TOPICS IN SEBASTIAN THRUN'S TOPIC PROFILE

Discipline	Probability
robotics	0.0728
computer science, artificial intelligence	0.0044
medicine, general & internal	0.0035
automation & control systems	0.0031
mining & mineral processing	0.0030
computer science, information systems	0.0018
communication	0.0012
information science & library science	0.0010
computer science, theory & methods	0.0006
behavioral sciences	0.0004

are merged together into a set of keywords P and mapped to these disciplines. For every discipline $d \in D$, we estimate the probability that the set of publications with description P belongs to d as:

$$\begin{aligned} \Pr(d|P) &= \frac{1}{|d|} \sum_{w \in d} \frac{\Pr(w|P)}{f(w,D)} \\ &\approx \frac{1}{|d|} \sum_{w \in d} \frac{f(w,P)}{|P| \cdot f(w,D)} \end{aligned} \quad (1)$$

where $f(w, P)$ is the number of occurrences of the discipline keyword w in P and $f(w, D)$ is the number of disciplines that contain w, which is used to measure the generality of that word. For example, in the discipline "computer science," "science" is more general than "computer" and therefore less important.

After estimating the probabilities of an author's publications over all disciplines, we can derive an author's topic profile. As an illustration, Table II shows the top 10 discipline topics in the profile of an author. Generally, the top topic d_1 is the author's main research area inferred from the publications, *robotics* in our example. We want to measure the probability that the author's publications belong to related disciplines. Our intuition is that topics related to the main research area contribute to the consistency of the profile. We therefore sum the probabilities of all related topics and normalized by the sum over all profile topics. Formally:

$$PTC = \frac{\sum_{i=1}^{N} \Pr(d_i|P)\delta(\sigma(d_1, d_i))}{\sum_{i=1}^{N} \Pr(d_i|P)} \quad (2)$$

where $\delta(\sigma(d_1, d_i))$ is the step function, equal to one if the similarity between d_1 and d_i is greater than zero, zero otherwise. N is the number of topics in the profile. This feature considers the interdisciplinary research collaborations of an author. The higher the value, the more likely that all the publications belong to consistent topics and therefore to the same author.

TABLE III
PERFORMANCE BASED ON VARIOUS FEATURES

Feature	Accuracy	F_1	AUC
CNV_1	0.57	0.54	0.65
CNV_2	0.67	0.63	0.74
CNV_3	0.70	0.66	0.74
PTC	0.73	0.73	0.79

IV. EVALUATION

In this section we evaluate the performance of classifiers that use the features described in § III to detect ambiguous author names. We study the proposed features separately and in combination. Following preliminary experiments with various classifiers, we present here results obtained with a simple logistic regression algorithm [16], which outperformed other methods in most cases. For each combination of features, we report on three performance measures: accuracy, area under ROC curve (AUC), and F_1. The AUC presents a tradeoff between true positives and false positives, while F_1 combines precision and recall. Average values of these measures are obtained by performing 10-fold cross-validation.

To train and test our classifiers, we manually labeled 500 author names as ambiguous or not. The names were selected among the top authors (ranked by H-index) from each of the top 100 disciplines (ranked by number of authors) in the Scholarometer database. We examined the publications retrieved from Google Scholar for each queried name. Titles, coauthors, venues, affiliations, and crowdsourced tags were inspected to obtain the ambiguity labels. Among the 500 author names, 283 are labelled not ambiguous and 217 ambiguous, which is a fairly balanced dataset.

Table III shows the results of two classes of features. Based on the the citations per name variation heuristic, we compare the accuracy of the classifier based on the percentage of citations accounted by top n ($n = 1, 2, 3$) name variations. The results suggest that using the top three name variations results in better detection of ambiguous names.

As a single feature, *PTC* achieves a relatively high accuracy, demonstrating that it is reasonable to consider the topic level

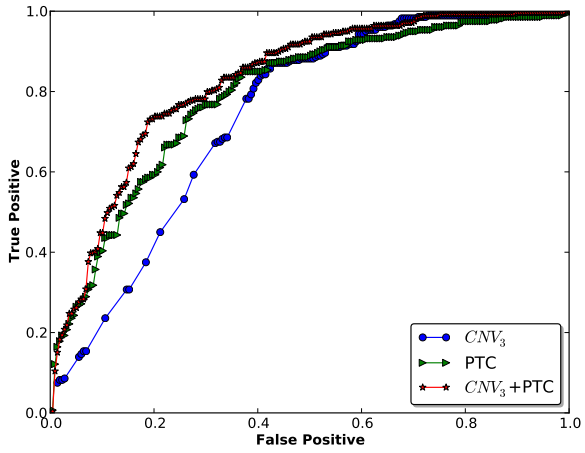

Fig. 1. ROC curves for 2 individual features and their combination. The classifier that combines the 2 features achieves an accuracy of 0.75, with $F_1 = 0.75$ and $AUC = 0.82$.

similarity of an author's publications. *PTC* benefits from the consideration of interdisciplinary collaborations.

Figure 1 summarizes the performance of two single features and their combination. The overall performance is quite promising, with accuracy around 75%.

V. Conclusions

We investigated the detection of ambiguous crowdsourced names in a social citation analysis system. Two classes of features were proposed. The first is a heuristic based on the percentage of citation accrued by the top name variations for an author. The second feature class utilizes crowdsourced data to detect ambiguity at the topic level. Our experiments show that the features work fairly well and yield accuracies around 70–73% when we classify using just a single feature. By combining the features, the accuracy of ambiguous author name detection goes up to 75%. In summary, the approach proposed here provides a promising first step toward the detection of ambiguous author names at query time.

In the future we plan to enhance the approach by exploring additional features, such as publication metadata, to further improve the accuracy of ambiguous author detection. Furthermore, for ambiguous names that enter into our system, we will explore the use of more traditional name disambiguation algorithms to partition publications into coherent clusters, combined with crowdsourcing techniques to let users select, merge, and/or split profiles for matching queried authors. The impact measures will be updated accordingly.

Acknowledgments

The work presented in this paper was performed while Xiaoling Sun and Lino Possamai were visiting the Center for Complex Networks and Systems Research (CNetS) at the Indiana University School of Informatics and Computing. We are deeply grateful to Diep Thi Hoang, without whom Scholarometer would not exist, for her continuous support. Thanks also to Snehal Patil, Heather Roinestad, and all the members of the Networks and agents Network (NaN) for helpful suggestions and discussions. We acknowledge support from Hongfei Lin at Dalian University of Technology, the China Scholarship Council, Massimo Marchiori at University of Padua, the University of Bologna, CNetS, and NSF (award IIS-0811994) for funding the computing infrastructure that hosts the Scholarometer service.

References

[1] H. Han, W. Xu, H. Zha, and C. Giles, "A hierarchical naive bayes mixture model for name disambiguation in author citations," in *Proc. of Symposium on Applied Computing*. ACM, 2005, pp. 1065–1069.

[2] D. Hoang, J. Kaur, and F. Menczer, "Crowdsourcing scholarly data," in *Proc. of Web Science Conference: Extending the Frontiers of Society On-Line (WebSci)*, 2010. [Online]. Available: http://journal.webscience.org/321/

[3] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, 2007.

[4] R. Cota, M. Gonçalves, and A. Laender, "A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries," in *Proc. of Simpnsio Brasileiro de Banco de Dados*, 2007, pp. 20–34.

[5] H. Han, C. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations," in *Proc. of International Conference on Digital Libraries*. ACM, 2004, pp. 296–305.

[6] H. Han, H. Zha, and C. Giles, "Name disambiguation in author citations using a k-way spectral clustering method," in *Proc. of International Conference on Digital Libraries*. ACM, 2005, pp. 334–343.

[7] J. Huang, S. Ertekin, and C. Giles, "Efficient name disambiguation for large-scale databases," *Knowledge Discovery in Databases: PKDD 2006*, pp. 536–544, 2006.

[8] B. Malin, "Unsupervised name disambiguation via social network similarity," in *Workshop on Link Analysis, Counterterrorism, and Security*, vol. 1401, 2005, pp. 93–102.

[9] J. Soler, "Separating the articles of authors with the same name." *Scientometrics*, vol. 72, no. 2, pp. 281–290, 2007.

[10] Y. Song, J. Huang, I. Councill, J. Li, and C. Giles, "Efficient topic-based unsupervised name disambiguation," in *Proc. of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2007, pp. 342–351.

[11] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, and J.-M. Ho, "Author name disambiguation for citations using topic and web correlation," *Research and Advanced Technology for Digital Libraries*, pp. 185–196, 2008.

[12] H. Al-Mubaid and P. Chen, "Biomedical term disambiguation: An application to gene-protein name disambiguation," in *Proc. of International Conference of Information Theory: New Generations*, 2006, pp. 606–612.

[13] D. Smith and G. Crane, "Disambiguating geographic names in a historical digital library," in *Proc. of European Conference on Digital Libraries*. Springer-Verlag, 2001, pp. 127–136.

[14] G. Mann and D. Yarowsky, "Unsupervised personal name disambiguation," in *Proc. of the seventh conference on Natural language learning at HLT-NAACL 2003*. Association for Computational Linguistics, 2003, pp. 33–40.

[15] A. Culotta, P. Kanani, R. Hall, M. Wick, and A. McCallum, "Author disambiguation using error-driven machine learning with a ranking loss function," in *Sixth International Workshop on Information Integration on the Web*, 2007.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.