Information Processing and Management 49 (2013) 454–464

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Ambiguous author query detection using crowdsourced digital library annotations

Xiaoling Sun [a,b,*], Jasleen Kaur [b], Lino Possamai [c,b], Filippo Menczer [b]

[a] Dept. of Computer Science and Technology, Dalian University of Technology, China
[b] School of Informatics and Computing, Indiana University, Bloomington, USA
[c] Department of Pure and Applied Mathematics, University of Padua, Italy

ARTICLE INFO

*Article history:*
Received 9 February 2012
Received in revised form 30 August 2012
Accepted 11 September 2012
Available online 15 October 2012

*Keywords:*
Ambiguous name detection
Data mining
Citation analysis
Scholarly data
Discipline annotations

ABSTRACT

The name ambiguity problem is especially challenging in the field of bibliographic digital libraries. The problem is amplified when names are collected from heterogeneous sources. This is the case in the *Scholarometer* system, which performs bibliometric analysis by cross-correlating author names in user queries with those retrieved from digital libraries. The uncontrolled nature of user-generated annotations is very valuable, but creates the need to detect ambiguous names. Our goal is to detect ambiguous names at query time by mining digital library annotation data, thereby decreasing noise in the bibliometric analysis. We explore three kinds of heuristic features based on citations, metadata, and crowdsourced topics in a supervised learning framework. The proposed approach achieves almost 80% accuracy. Finally, we compare the performance of ambiguous author detection in Scholarometer using Google Scholar against a baseline based on Microsoft Academic Search.

## 1. Introduction

Bibliometric methods measure the impact of papers, researchers, journals, and even organizations. Many measures based on citation data have been proposed to estimate quantitatively the impact of an author. Examples include the H-index (Hirsch, 2005) and the G-index (Egghe, 2006). These measures rely on citations; therefore, consistent, accurate, and up-to-date citation data is critical for an accurate assessment of author impact.

The publications of an author are typically identified by the author's name. However, we cannot always correctly match publications with authors because names can be ambiguous. There are two types of name ambiguity: multiple name variations and same-named authors (Han, Xu, Zha, & Giles, 2005). Here we deal with the latter, which makes citation-based impact analysis noisy and therefore complicates efforts to measure impact.

*Scholarometer* (`scholarometer.indiana.edu`) is a social tool that helps evaluate the impact of authors (Hoang, Kaur, & Menczer, 2010; Kaur et al., 2012). In our approach to scholarly citation analysis, information that is crowdsourced from end-users of the system forms the very basis for the service provided. To decrease the noise in the citation analysis, we want to detect ambiguous names at query time.

Prior work in name disambiguation is based on 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 detection of ambiguous names. When an ambiguous name is detected in this

[*] Corresponding author. Address: 919 E. 10th St., Bloomington, IN 47408, USA. Tel.: +1 812 327 4576; fax: +1 812 855 0600.
*E-mail address:* xlsun@mail.dlut.edu.cn (X. Sun).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.09.001

setting the tool should prompt the user to refine the query, e.g., by adding keywords to make the query more specific. We adopt this definition of ''ambiguity'' based on user queries; thus our problem formulation: *Given a set of publications obtained by querying a digital library, decide if the author names of these publications match the name in the query*.

### *1.1. Contributions and outline*

After background on related work in Section 2 and description of data crowdsourced through Scholarometer in Section 3, we present the proposed approach in Section 4, which includes the following contributions:

- A heuristic based on name variations and citations (Section 4.1).
- A two-step method to capture the consistency between coauthor, title, and venue metadata across publications (Section 4.2).
- An algorithm to measure the coherence between the topics associated with title and venue metadata (Section 4.3.1).
- An algorithm to measure the consistency between topics associated with publication metadata with the addition of crowdsourced discipline annotations for authors (Section 4.3.2).

In Section 5 we evaluate these features in a supervised learning setting and show the effectiveness of combining them with each other. We also show that our approach outperforms a baseline derived from Microsoft Academic Search.

## 2. Related work

The problem of ambiguous names is important because it affects the quality of content and services in digital libraries. For example, if we want to evaluate the impact of author ''J. Smith'' using Google Scholar, we can see that the publications belong to several different authors with the same name. If all these publications were used, the impact of ''J. Smith'' would be severely overestimated. There is little data in the literature on the prevalence of ambiguous names. Kang et al. (2009) found an average of 1.71 distinct individuals per author name in Korean publications. Our own analysis based on Scholarometer data suggests that 25% of the queried author names are ambiguous.

The literature on disambiguation is mainly categorized into supervised and unsupervised learning approaches. Supervised learning approaches (Culotta, Kanani, Hall, Wick, & McCallum, 2007; Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004; Huang, Ertekin, & Giles, 2006) 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. However, it is expensive for humans to label sufficiently many names for training in very large-scale digital libraries. One solution is a hybrid of manual and automatic methods, as in the paradigm of active learning (Kanani, McCallum, & Pal, 2007), which can iteratively detect the most informative examples for manual curation. Recently, Levin, Krawzyk, Bethard, and Jurafsky (2012) proposed a self-supervised algorithm for author disambiguation in large bibliographic databases. Veloso, Ferreira, Gonçalves, Laender, and Meira (2012) also introduced an associative author name disambiguation approach with self-training capabilities.

Unsupervised approaches (Bhattacharya and Getoor, 2007; Cota et al., 2007; Han, Xu, et al., 2005; Han, Zha, et al., 2005; Malin, 2005; Soler, 2007; Song et al., 2007; Yang et al., 2008) do not use training examples; instead, they exploit publication features to merge similar publications into clusters, such as coauthorship (Kang et al., 2009). In general, supervised approaches perform better because they are tuned specifically to determine the relative importance of, and interactions among, different features of the data, such as coauthors, venues, titles, and affiliations.

Recently, *Microsoft Academic Search* (`academic.research.microsoft.com`) and *Google Scholar* (`scholar.google.com`) have introduced hybrid approaches that combine automatic clustering with manual curation, a crowdsourcing approach to name disambiguation. This approach is not applicable in our setting because the Scholarometer system does not require authors to create profiles.

The approach taken by the Scholarometer system for disambiguation is based on supervised learning, but defines the 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. Our first attempt to deal with ambiguous author names deployed a simple heuristic based on name variations and citations (Hoang et al., 2010) discussed in Section 4.1. We improved on the simple heuristic by analyzing the topic-level consistency of author publications (Sun, Kaur, Possamai, & Menczer, 2011), which is mentioned in Section 4.3.1. With increased popularity the number of authors in Scholarometer has grown significantly, revealing many undetected ambiguous names. This has motivated us to explore more features for better detection of ambiguous names.

## 3. Crowdsourced data

In this section we outline the data acquisition in the Scholarometer system and the statistics of the crowdsourced data. Further details can be found elsewhere (Kaur et al., 2012).

### 3.1. Data acquisition

Crowdsourcing is an approach to harness knowledge from a community via Web platforms in order to solve practical problems. Scholarometer applies crowdsourcing to scholarly annotations. Users provide disciplinary annotations in exchange for access to citation data obtained from querying bibliographic services. Therefore, we consider two sources of data: (i) citation data from an online bibliographic digital library and (ii) user-supplied annotations of authors with discipline tags.

As a browser extension, Scholarometer accepts queries about authors, which must include discipline annotations. The tool performs citation analysis, fetching data from Google Scholar on behalf of the user. Fig. 1 shows a screenshot with query results. The data that we collect comes from users, so it is naturally noisy. A blacklist is employed to prevent spammers from polluting databases. We also apply manual and automatic data cleaning techniques to deal with noisy annotations (tags). Another important issue is the ambiguity of author names, which is the topic of this paper.

### 3.2. Data analysis

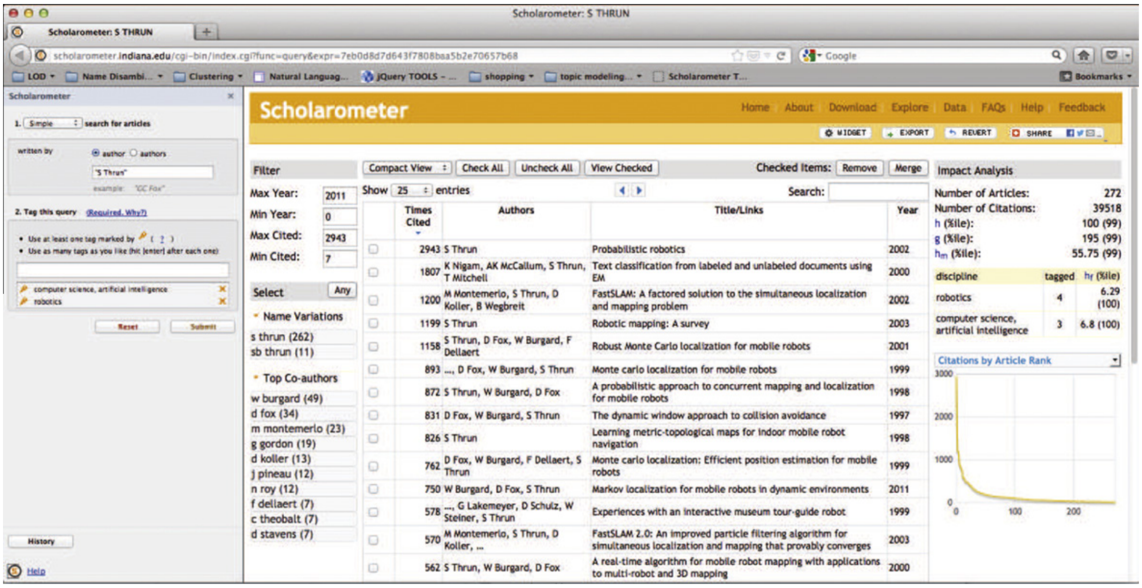
The Scholarometer system was first released in November 2009. At the time of writing, the database has collected about 1.9 million articles by 26 thousand authors in 1200 disciplines. Fig. 2 displays the top 15 discipline tags based on the number of authors. The Scholarometer database was initially dominated by computing-related disciplines due to the publicity received by the tool in the computer and information science communities. Disciplinary coverage has since grown. Further statistics for authors and disciplines are presented by Kaur et al. (2012) and available on the Scholarometer website (http://www.scholarometer.indiana.edu/explore.html).

## 4. Ambiguous name detection

Our algorithm extracts features from all publications retrieved for the queried author name and performs binary classification to estimate the likelihood that the set of publications belongs to the same author. In this section we describe three classes of features of the publications of an author.

### 4.1. Name variations and citations

Typical author names have two or three variations. To extract name variations for a queried author, we compute a heuristic similarity measure between the author names from the retrieved publications and the queried name. All the names with similarity above an empirical threshold are considered as variations and ranked by the citations counts of the corresponding papers. For example, the name variations “s thrun” and “sb thrun” are shown in Fig. 1. We then 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



Scholarometer

1. Simple search for articles

written by ● author ○ authors

"S Thrun"

example: "GC Fox"

2. Tag this query (Required. Why?)

- Use at least one tag marked by ( ? )
- Use as many tags as you like (hit [enter] after each one)

computer science, artificial intelligence

robotics

Filter

Max Year: 2011

Min Year: 0

Max Cited: 2943

Min Cited: 7

Select

Name Variations

s thrun (262)

sb thrun (11)

Top Co-authors

w burgard (49)

d fox (34)

m montemerlo (23)

g gordon (19)

d koller (13)

j pineau (12)

n roy (12)

f dellaert (7)

c theobalt (7)

d stavens (7)

Checked Items:

Show 25 entries

Search:

| | Times Cited | Authors | Title/Links | Year |
|---|---|---|---|---|
| □ | 2943 | S Thrun | Probabilistic robotics | 2002 |
| □ | 1807 | K Nigam, AK McCallum, S Thrun, T Mitchell | Text classification from labeled and unlabeled documents using EM | 2000 |
| □ | 1200 | M Montemerlo, S Thrun, D Koller, B Wegbreit | FastSLAM: A factored solution to the simultaneous localization and mapping problem | 2002 |
| □ | 1199 | S Thrun | Robotic mapping: A survey | 2003 |
| □ | 1158 | S Thrun, D Fox, W Burgard, F Dellaert | Robust Monte Carlo localization for mobile robots | 2001 |
| □ | 893 | ..., D Fox, W Burgard, S Thrun | Monte carlo localization for mobile robots | 1999 |
| □ | 872 | S Thrun, W Burgard, D Fox | A probabilistic approach to concurrent mapping and localization for mobile robots | 1998 |
| □ | 831 | D Fox, W Burgard, S Thrun | The dynamic window approach to collision avoidance | 1997 |
| □ | 826 | S Thrun | Learning metric-topological maps for indoor mobile robot navigation | 1998 |
| □ | 762 | D Fox, W Burgard, F Dellaert, S Thrun | Monte carlo localization: Efficient position estimation for mobile robots | 1999 |
| □ | 750 | W Burgard, D Fox, S Thrun | Markov localization for mobile robots in dynamic environments | 2011 |
| □ | 578 | ..., G Lakemeyer, D Schulz, W Steiner, S Thrun | Experiences with an interactive museum tour-guide robot | 1999 |
| □ | 570 | M Montemerlo, S Thrun, D Koller, ... | FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges | 2003 |
| □ | 562 | S Thrun, W Burgard, D Fox | A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping | 2000 |

Impact Analysis

| | |
|---|---|
| Number of Articles: | 272 |
| Number of Citations: | 39518 |
| h (%ile): | 100 (99) |
| g (%ile): | 195 (99) |
| $h_m$ (%ile): | 55.75 (99) |

| discipline | tagged | $h_f$ (%ile) |
|---|---|---|
| robotics | 4 | 6.29 (100) |
| computer science, artificial intelligence | 3 | 6.8 (100) |

**Fig. 1.** Illustration of the Scholarometer interface. The browser extension interface (left) lets users enter queries and tags. The main browser window (center) is for presenting and manipulating bibliographic data. Citation analysis results are shown on the right.

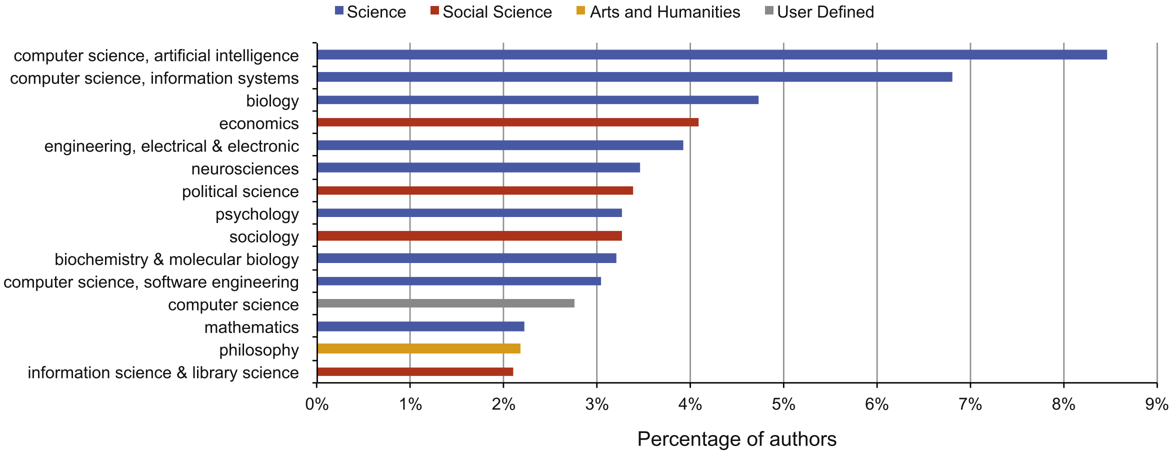


**Fig. 2.** Percentage of authors tagged with 15 most common disciplines. Note that the sets of authors in these disciplines may overlap, as authors are often tagged with multiple disciplines.

on impact measures. If the top name variations account for a low fraction of the citations, it is reasonable to assume the contrary.

The first iteration of this heuristic used in Scholarometer had a fixed threshold (90%) for the fraction of citations to papers corresponding to the *top three* name variations (Hoang et al., 2010). 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 *citations per name variation* ($CNV_n$).

### 4.2. Metadata consistency

Generally, an author is likely to collaborate with a certain group of colleagues, write papers on similar topics, and publish papers in similar journals or conferences. The metadata associated with these publications by the same author should be consistent; inconsistencies between publication patterns are evidence of an ambiguous author name.

We explore three common publication attributes: coauthors, title, and venue. For each attribute we compute the average similarity over pairs of publications. Some preprocessing, including stopword removal and stemming, takes place before computing any similarity.

Two kinds of similarity methods are used: cosine similarity and overlap coefficient. Cosine similarity is defined as:

$$\sigma(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vector representations (Salton & Buckley, 1988). We use simple term frequencies as vector weights. The overlap coefficient is defined as:

$$overlap(x,y) = \frac{|x \cap y|}{\min(|x|,|y|)} \tag{2}$$

where $x$ and $y$ are two sets of binary features.

For title similarity we treat each title as a vector of words and use cosine similarity. For venue and coauthor similarity we compute the overlap coefficient between every pair of venues or coauthors. We call these metadata features $MD_{title}$, $MD_{venue}$, and $MD_{coauthor}$.

There are two issues related to the above features. First, titles and venues of publications are usually very short. Therefore, the similarity scores over pairs of titles or venues are often very small or zero. Second, the comparisons among all pairs of publications are computationally costly for a query-time algorithm. To alleviate these problems, we propose a two-stage strategy:

1. *Coauthor clustering:* The clustering algorithm follows the basic assumption that if two papers have at least one coauthor in common, they belong to the same author (Cota et al., 2007). This algorithm, though simple, groups some publications together with high confidence. In every cluster, the titles and venues are merged together.
2. *Pairwise similarity:* Based on the first step, we calculate the title and venue similarities over pairs of clusters in a similar way to $MD_{title}$ and $MD_{venue}$.

We call these features $MD_{coauthor+title}$ and $MD_{coauthor+venue}$. This two-stage strategy alleviates the sparsity problem and reduces the computational cost of the similarity computation, even though these features are still expensive to compute for query-time detection. Therefore, in Section 5.2 we propose a sampling strategy to limit the number of cluster comparisons.

### 4.3. Topic consistency

We leverage the discipline tags crowdsourced from the users of the Scholarometer system (Fig. 1) to capture topical consistency between authors and publications. Each discipline can be represented by a vector of authors who have been tagged with it. The vector weight is the number of times that an author has been tagged with a discipline. The cosine similarity $\sigma$ between two disciplines, calculated by Eq. (1), reflects the strength of cross-disciplinary collaborations between authors in these two disciplines. As an illustration, we list eight pairs of highly related disciplines in Table 1.

#### 4.3.1. Publication topic consistency

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, in Fig. 3 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*).

We use a subset of the crowdsourced tags from a controlled vocabulary, namely the JCR categories. These 242 preexisting disciplines are composed of *Science Citation Index Expanded*, *Social Sciences Citation Index*, and *Arts and Humanities Citation Index* from Thomson Reuters' *Web of Science*.

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

$$\Pr(d|P) = \frac{1}{|d|}\sum_{w\in d}\frac{\Pr(w|P)}{f(w,D)} \approx \frac{1}{|d|}\sum_{w\in d}\frac{f(w,P)}{|P|\cdot f(w,D)} \tag{3}$$

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.

As an illustration, Table 2 shows the top five discipline topics in the profiles of three authors. Generally, the top one topic is the author's main research area as inferred from the publications. 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 normalize by the sum over all profile topics. Formally:

$$PTC = \frac{\sum_{i=1}^{N}\Pr(d_i|P)\delta(\sigma(d_1,d_i))}{\sum_{i=1}^{N}\Pr(d_i|P)} \tag{4}$$

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, and zero otherwise. $N$ is the number of topics in the profile. This feature considers the interdisciplinary research collaborations of an author.

#### 4.3.2. Author-publication topic consistency

Suppose that the majority of the publications of an author have high metadata and publication topic consistency, but the crowdsourced tags are inconsistent with the publication topics. For example, an author is tagged with "chemistry", however, the publications appear to be consistently related to "literature". Since there is low similarity between these two disciplines, we may infer that the author name is ambiguous.

Based on the above scenario, we propose a new feature called *author-publication topic consistency* (*APTC*). Let us define it as the similarity between the publications profile and the crowdsourced discipline tags:

$$APTC = \max_{d_i\in T_A, d_j\in T_P}\sigma(d_i,d_j) \tag{5}$$

**Table 1**
Eight highly related discipline pairs and their cosine similarities.

| | | |
|---|---|---|
| Materials science, multidisciplinary | Optics | 0.82 |
| Neurosciences | Language and linguistics | 0.78 |
| Surgery | Transplantation | 0.75 |
| Geology | Paleontology | 0.73 |
| Behavioral sciences | Biodiversity conservation | 0.70 |
| Demography | Management | 0.70 |
| Anatomy and morphology | Obstetrics and gynecology | 0.63 |
| Ecology | Paleontology | 0.63 |

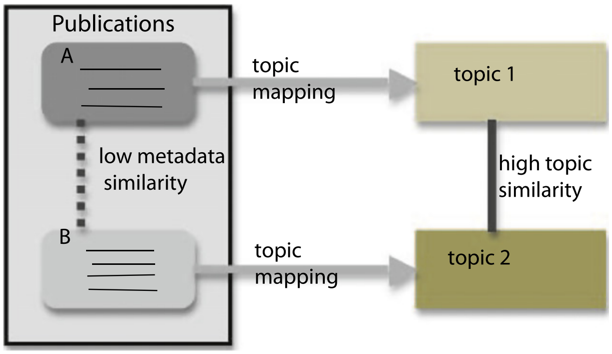


**Fig. 3.** Metadata vs. topic similarity.

**Table 2**
Top five disciplines in the topic profiles of three authors. The topic probabilities are inferred from the author's publications. Note that since the profiles only rely on title and venue keywords, they can be mapped to incorrect disciplines.

| Discipline | Probability |
|---|---|
| *Sebastian Thrun* | |
| Robotics | 0.0728 |
| Computer science, artificial intelligence | 0.0044 |
| Medicine, general and internal | 0.0035 |
| Automation and control systems | 0.0031 |
| Mining and mineral processing | 0.0030 |
| *Sigmund Freud* | |
| Psychology, psychoanalysis | 0.0057 |
| Social work | 0.0050 |
| Engineering, civil | 0.0027 |
| Psychology | 0.0023 |
| Literary theory and criticism | 0.0018 |
| *Robert M May* | |
| Ecology | 0.0348 |
| Infectious diseases | 0.0076 |
| Psychology, biological | 0.0049 |
| Parasitology | 0.0038 |
| Evolutionary biology | 0.0022 |

where $T_A$ is the set of crowdsourced tags for the author and $T_P$ is set of topics for the publications. For the latter we collect disciplines that contribute to the numerator in Eq. (4).

This feature considers user-generated discipline annotations. Since reliability of data is an issue in the computation of the *ATPC*, we only wish to include reliable tags in $T_A$. We view each query as a vote for the discipline tags of the queried author. The number of votes together with the number of tags are used to determine heuristically which tags are reliable for each author (Kaur et al., 2012).

## 5. Evaluation

Here we present results obtained with a simple logistic regression algorithm from Weka (Hall et al., 2009), which outperformed other methods in most cases. For each combination of features, we report on three performance measures: accuracy (Acc), area under ROC curve (AUC), and $F_1$. Average values of these measures are obtained by performing 10-fold cross-validation.

To train and test our classifier, we manually labeled 500 author names. 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. Four graduate students 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. We randomly selected 250 authors and judged them twice, with excellent inter-rater agreement (Cohen's $\kappa = 0.84$). The other 250 authors were judged once. Finally, among the 500 author names, 283 were labeled "not ambiguous" and 217 "ambiguous."

### 5.1. Name variations and citations

Based on the citations-per-name-variation heuristic, we compare the accuracy of the classifier based on the percentage of citations accounted for by the top $n$ ($n$ = 1, 2, 3) name variations. The ROC curves in Fig. 4 and measures in Table 3 suggest that using the top three name variations results in better detection of ambiguous names.

We notice that when there are more than three name variations, $CNV_3$ is a good feature, which can be seen from the upper right part of the curves in Fig. 4. For the authors with less than three name variations, other features are necessary for detecting ambiguity. Based on the accuracy, we select $CNV_3$.

### 5.2. Metadata consistency

To compute the metadata features efficiently, we exploit estimated knowledge of author impact. We only consider the top $h$ (author's H-index) publications, as they are the ones that affect the impact computation. While this choice precludes us from detecting certain ambiguities, namely those that do not affect the impact measure, it is necessary in order to deal with the large amount of noise present in the tail of the publications returned by Google Scholar.

Fig. 5 shows the accuracy achieved with metadata features computed over samples of publication/cluster pairs of different sizes. A sample of 200 pairs provides an adequate balance between efficiency and accuracy. We use this sample size to produce the results shown in Tables 4 and 5.

Table 4 shows the results of the combinations of the metadata features. As expected, title similarity $MD_{title}$ performs better than the other two features. Venues have relatively fewer words, resulting in smaller overlap. An author may collaborate with different groups of people at different times, so the average coauthor similarity may be low even for unambiguous authors. Table 5 shows the results of our two-stage strategy. From both tables, we see that the best results are obtained by combining the two metadata consistency features with coauthor clustering.

### 5.3. Topic consistency and summary

Table 6 shows that $PTC$ achieves relatively high accuracy as a single feature, demonstrating that it is reasonable to consider the topic-level similarity of an author's publications. $PTC$ benefits from the consideration of interdisciplinary collaborations. $APTC$ is not as good as $PTC$; inconsistencies between publication topics and author tags may be caused by poor tag choices rather than ambiguous names. In combination with all the other four features, however, $APTC$ does provide a slight advantage (Table 7).

Fig. 6 and Table 7 summarize the performance of five single features and their combination. The performance of the best classifier, combining all features, is quite promising, with an accuracy of 79%. This result improves upon the 75% accuracy achieved with the combined feature $CNV_3 + PTC$ (Sun et al., 2011).

The above results, while promising, are based on a relatively small hand-curated dataset of 500 authors. To evaluate our algorithm on a larger dataset, we can follow the semi-supervised learning approach of Levin et al. (2012), and apply our best classifier to the entire set of authors in our database. 4202 authors (25% at the time) were labeled "ambiguous." To obtain a conservative estimate of classification accuracy, we drew a random sample of 4279 authors labeled "not ambiguous," yielding a balanced dataset. We then ran 10-fold cross-validation on the resulting set of 8481 authors, obtaining an accuracy of 92.3%.

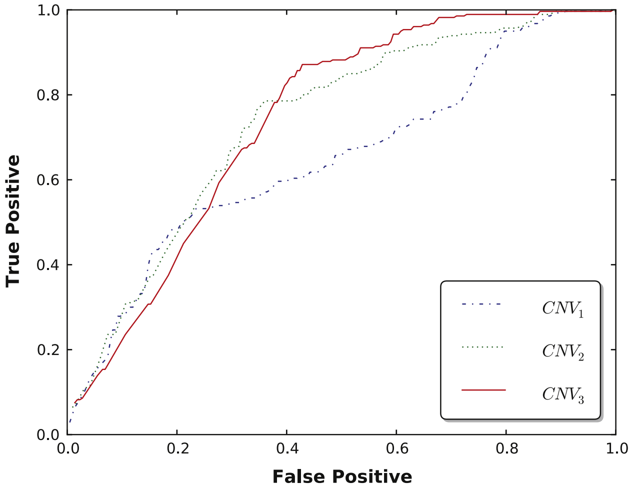
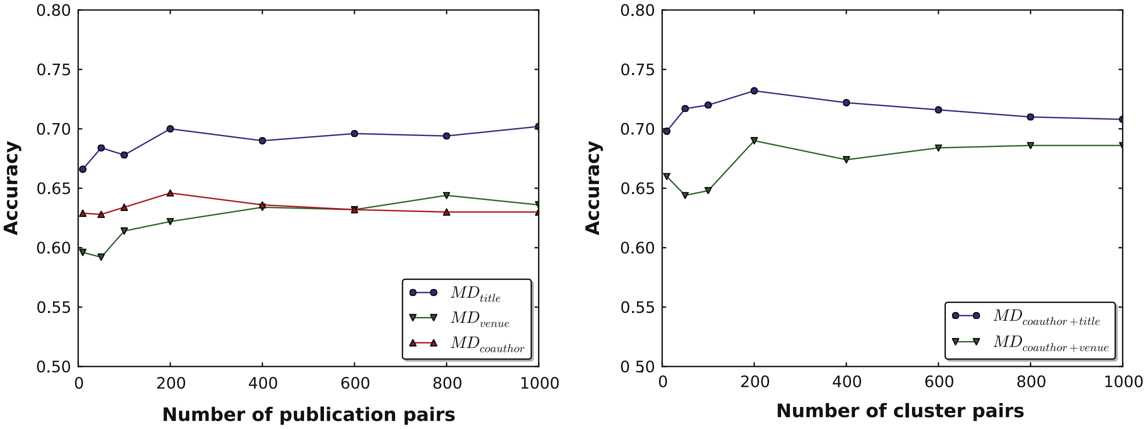


**Fig. 4.** ROC curves based on citations per name variation features.

**Table 3**
Performance based on citation per name variation (*CNV*) heuristics.

| Feature | Acc | $F_1$ | 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 |



(a) Publications

(b) Clusters

**Fig. 5.** Accuracy vs. number of pairs in sample.

**Table 4**
Performance based on metadata without clustering.

| Feature | Acc | $F_1$ | AUC |
|---|---|---|---|
| $MD_{title}$ | 0.70 | 0.70 | 0.75 |
| $MD_{venue}$ | 0.62 | 0.62 | 0.66 |
| $MD_{coauthor}$ | 0.65 | 0.65 | 0.68 |
| $MD_{title}$, $MD_{coauthor}$ | 0.72 | 0.72 | 0.79 |
| $MD_{venue}$, $MD_{coauthor}$ | 0.71 | 0.71 | 0.77 |
| Three metadata features | 0.72 | 0.72 | 0.77 |

**Table 5**
Performance based on metadata with clustering.

| Feature | Acc | $F_1$ | AUC |
|---|---|---|---|
| $MD_{coauthor+title}$ | 0.73 | 0.73 | 0.77 |
| $MD_{coauthor+venue}$ | 0.69 | 0.69 | 0.70 |
| $MD_{coauthor+title}$, $MD_{coauthor+venue}$ | 0.74 | 0.74 | 0.80 |

**Table 6**
Performance based on topic consistency.

| Feature | Acc | $F_1$ | AUC |
|---|---|---|---|
| *PTC* | 0.73 | 0.73 | 0.79 |
| *APTC* | 0.64 | 0.64 | 0.70 |

### *5.4. Comparison with a baseline*

For better assessment it is desirable to compare our approach with a suitable baseline. However, we are not aware of other methods in the literature that perform ambiguous author name detection at query time as we have defined the task here.

**Table 7**
Summary performance based on various features.

| Feature | Acc | $F_1$ | AUC |
|---|---|---|---|
| $CNV_3$ | 0.70 | 0.66 | 0.74 |
| $MD_{coauthor+title}$ | 0.73 | 0.73 | 0.77 |
| $MD_{coauthor+venue}$ | 0.69 | 0.69 | 0.70 |
| *PTC* | 0.73 | 0.73 | 0.79 |
| *APTC* | 0.64 | 0.64 | 0.70 |
| $CNV_3$ + *PTC* | 0.75 | 0.75 | 0.82 |
| Four features (without *APTC*) | 0.78 | 0.78 | 0.84 |
| Best (all features combined) | 0.79 | 0.79 | 0.84 |

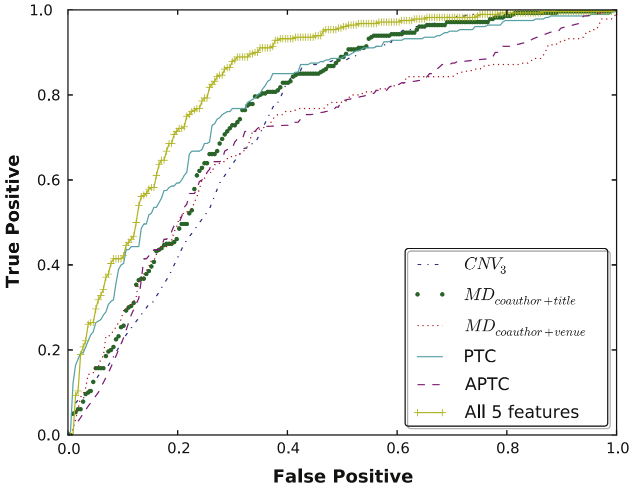


**Fig. 6.** ROC curves for five individual features and their combination.

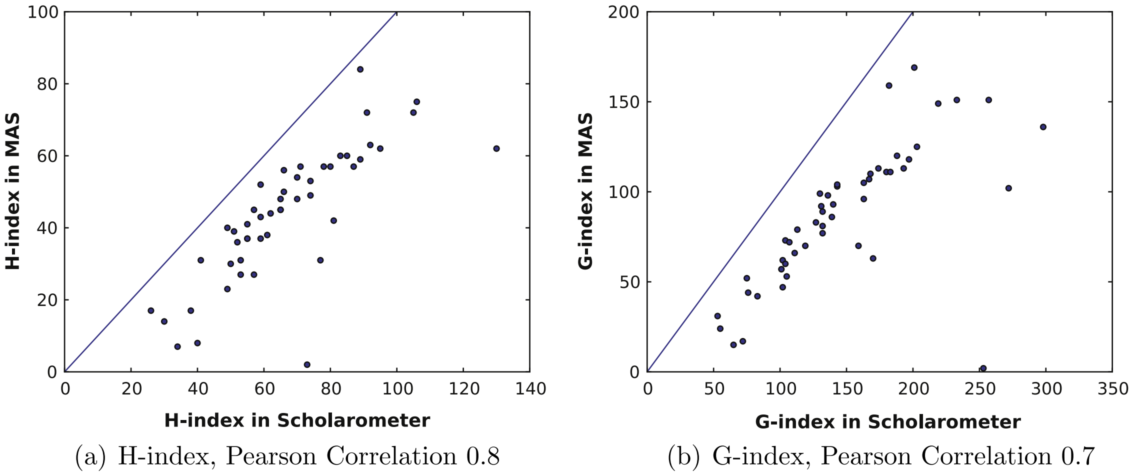
In search of a baseline for our evaluation, let us consider *Microsoft Academic Search* (*MAS*). MAS provides name disambiguation by clustering the papers associated with an author name into profiles. To derive a suitable baseline from MAS, we must interpret its output as if it were a binary ambiguous name detection system. Naturally, this is not what MAS is designed to do, but it is nevertheless interesting to compare the performance of our ambiguous name detection system with a baseline derived from the state of the art.

A number of measures must be taken to ensure that our comparison is not biased against the baseline and in favor of our approach. Our analysis suggests that Google Scholar has higher coverage than MAS in all fields. Yet, the publication coverage of MAS is expanding rapidly. At the time of this writing, MAS has relatively good coverage in the fields of computer science, chemistry, mathematics, engineering, and physics. To conduct a relatively fair comparison, we select a candidate test set of authors who are tagged reliably with these five disciplines. Among the 500 labeled authors, 184 meet this criterion. For each of these authors, we query MAS to make sure that the publications in MAS and Google Scholar (the Scholarometer data source) are comparable. In MAS an author may have several profiles. If the aggregate number of publications across the profiles is *h* (author's H-index) or more, we consider the coverage sufficient to carry out the comparison. After filtering out authors with an insufficient number of publications, we are left with a test dataset of 139 authors; 60 of these are ambiguous and 79 not ambiguous according to our definition.

For the sake of obtaining a baseline ambiguity detector, we ignore the profile details and focus solely on the number of profiles returned by MAS for every author name in the test set. A single profile is interpreted as a binary classification of "not ambiguous" and multiple profiles as an "ambiguous" name detection. To avoid a bias in favor of our system, our classifier is trained on the 361 excluded authors and then tested on the same set of 139 authors as MAS.

**Table 8**
Confusion matrices.

| | MAS | | Our system | |
|---|---|---|---|---|
| | Positive | Negative | Positive | Negative |
| Positive | 29 | 50 | 44 | 35 |
| Negative | 0 | 60 | 6 | 54 |



(a) H-index, Pearson Correlation 0.8 (b) G-index, Pearson Correlation 0.7

**Fig. 7.** Scatter plots of citation based impact measures for highly cited computer science authors in MAS and Scholarometer. Points would be aligned on the diagonal line if the two systems had perfectly consistent impact measures.

The confusion matrices generated by our detector and the MAS baseline are shown in Table 8. MAS has a high false negative rate of 63% compared to our system whose false negative rate is 44%. This is because MAS tends to split the publications of an author into small coherent groups so that an unambiguous author name may have several profiles. In the context of citation analysis this is likely to result in lower impact measures for authors. As an illustration, we compare the H-index and G-index values reported by the two systems for 49 highly cited computer scientists with unambiguous names in Fig. 7. While the values are correlated, MAS tends to report lower impact measures. For the same reason, MAS has no false positives, while our system misses an ambiguous author, overestimating impact of six authors.

Overall, the accuracy of our detector is 71% vs. 64% for the baseline. Note that our system's accuracy is lower than reported in Table 7; here we have a predefined test set and cannot use cross-validation. With all the caveats of the baseline comparison (small test set, differential coverage, and binary classification), we consider these results encouraging. Our approach achieves better performance for ambiguous author name detection in the context of citation analysis.

## 6. Conclusions

We investigated the detection of ambiguous crowdsourced names in a social citation analysis system. Three classes of features were explored, extending previous work. The first is a heuristic based on the percentage of citation accrued by the top name variations for an author. The second is based on consistency of metadata, including titles and venues across coauthor clusters. The third utilizes crowdsourced data to detect ambiguity at the topic level. Our experiments show that these features work fairly well and yield accuracies around 64–73% when we classify using just a single feature. By combining all the features, the accuracy of ambiguous author name detection increases to 79%. We also find that the accuracy of our approach compares favorably with a baseline derived from Microsoft Academic Search when testing on a subset of authors.

We have implemented the proposed method into the latest version of the Scholarometer system, and the detector is sufficiently efficient to be compatible with the real-time environment, which was one of our goals. On average, it takes 1.06 s to compute the features and run the detection algorithm once the data is obtained from Google Scholar.

In the future we plan to enhance the approach by exploring additional features, such as publication years, to further improve the accuracy of ambiguous author detection. Furthermore, for undetected ambiguous names, 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.

## 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.indiana.edu`) 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 Mohsen JafariAsbagh, Snehal Patil, Heather Roinestad, Stasa Milojevic, John McCurley, and all the members of the Networks and agents Network (`cnets.indiana.edu/groups/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, the Lilly Endowment, and NSF (Award IIS-0811994) for funding the computing infrastructure that hosts the Scholarometer service.

## References

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1), 5.

Cota, R., Gonçalves, M., & Laender, A. (2007). A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries. In *Proc. of brazilian symposium on databases (SBBD)* (pp. 20–34).

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. In *Sixth intl. workshop on information integration on the web*.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics, 69*(1), 131–152.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The weka data mining software: An update. *SIGKDD Explorations Newsletter, 11*(1), 10–18.

Han, H., Giles, C., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proc. of intl. conference on digital libraries* (pp. 296–305).

Han, H., Xu, W., Zha, H., & Giles, C. (2005). A hierarchical naive Bayes mixture model for name disambiguation in author citations. In *Proc. of symposium on applied computing* (pp. 1065–1069).

Han, H., Zha, H., & Giles, C. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In *Proc. of intl. conference on digital libraries* (pp. 334–343).

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569.

Hoang, D., Kaur, J., & Menczer, F. (2010). Crowdsourcing scholarly data. In *Proc. of web science conference: Extending the frontiers of society on-line (WebSci)*. <http://journal.webscience.org/321/>.

Huang, J., Ertekin, S., & Giles, C. (2006). Efficient name disambiguation for large-scale databases. *Knowledge Discovery in Databases: PKDD*, 536–544.

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the web. In *Proc. of the 20th intl. joint conference on artificial intelligence* (pp. 429–434).

Kang, I., Na, S., Lee, S., Jung, H., Kim, P., Sung, W., et al (2009). On co-authorship for author disambiguation. *Information Processing and Management, 45*(1), 84–97.

Kaur, J., Hoang, D. T., Sun, X., Possamai, L., JafariAsbagh, M., Patil, S., et al (2012). Scholarometer: A social framework for analyzing impact across disciplines. *PLoS One, 7*(9), e43235. http://dx.doi.org/10.1371/journal.pone.0043235.

Levin, M., Krawzyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*.

Malin, B. (2005). Unsupervised name disambiguation via social network similarity. In *Workshop on link analysis, counterterrorism, and security* (Vol. 1401, pp. 93–102).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*, 513–523.

Soler, J. (2007). Separating the articles of authors with the same name. *Scientometrics, 72*(2), 281–290.

Song, Y., Huang, J., Councill, I., Li, J., & Giles, C. (2007). Efficient topic-based unsupervised name disambiguation. In *Proc. of the 7th ACM/IEEE-CS joint conference on digital libraries* (pp. 342–351).

Sun, X., Kaur, J., Possamai, L., & Menczer, F. (2011). Detection ambiguous author names in crowdsourced scholarly data. In *Proc. of IEEE intl. conference on social computing* (pp. 568–571).

Veloso, A., Ferreira, A., Gonçalves, M., Laender, A., & Meira, W. (2012). Cost-effective on-demand associative author name disambiguation. *Information Processing and Management, 48*, 680–697.

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