The Knowledge Engineering Review, Vol. 32, e22, 1–24. © Cambridge University Press, 2017
doi:10.1017/S0269888917000182

A survey of author name disambiguation techniques: 2010–2016

IJAZ HUSSAIN and SOHAIL ASGHAR

Department of Computer Science, COMSATS Institute of Information Technology, Islamabad 45550, Pakistan;
e-mail: ijazhussain7979@hotmail.com, Sohail.Asg@gmail.com

Abstract

Digital libraries content and quality of services are badly affected by the author name ambiguity problem in the citations and it is considered as one of the hardest problems faced by the digital library researchers. Several techniques have been proposed in the literature for the author name ambiguity problem. In this paper, we reviewed some recently presented author name disambiguation techniques and give some challenges and future research directions. We analyze the recent advancements in this field and classify these techniques into supervised, unsupervised, semi-supervised, graph-based and heuristic-based techniques according to their problem formulation that is mainly used for the author name disambiguation. A few surveys have been conducted to review different techniques for the author name disambiguation. These surveys highlighted only the methodology adopted for author name disambiguation but did not critically review their shortcomings. This survey provides a detailed review of author name disambiguation techniques available in the literature, makes a comparison of these techniques at an abstract level and discusses their limitations.

1 Introduction

In this electronic era, digital libraries' (DLs) importance in academics is growing in leaps and bounds due to numerous factors such as cuts in budget for traditional libraries, nearly unlimited storage space at a much lower cost, ease of use, no physical boundary, round the clock availability and advances in information technology (Song *et al*., 2007; Palfrey, 2016; Weiss, 2016). DLs, for example, DBLP[1], MEDLINE[2], CiteSeer[3] arXiv[4], MAS[5], Google Scholar[6], and BDBComp[7] are being extensively used by the researchers to find scholarly literature for their research and discovery (Nicholson & Bennett, 2016).

In addition to the literature search facility, these DLs also provide some useful analysis and information functionality that is being used for better decision making by funding agencies and academic institutions for grants and individual's promotion decisions. It is presumed that these DLs contain and provide high-quality content to its users, however, they failed to provide (Lee *et al*., 2007). Christen (2006) and Ferreira *et al*. (2012) in their studies stated that the main sources of errors in DLs are the typographical, scanning and data conversion, find and replace, copy and paste, meta data, imperfect citation-gathering software, disparate citation formats, ambiguous author names, the decentralized generation of content (i.e., by

[1] http://dblp.uni-trier.de
[2] http://www.medline.com
[3] http://citeseerx.ist.psu.edu
[4] http://arxiv.org
[5] http://academic.research.microsoft.com
[6] http://scholar.google.com.pk
[7] http://www.lbd.dcc.ufmg.br/bdbcomp

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

means of automatic harvesting) and abbreviations of publication venue titles, etc. Among these sources of errors, a great attention is paid to the ambiguous author names from the research community due to its inherent difficulty. Specifically, name ambiguity arises when a set of citation records contains ambiguous author names and it may appear in two different forms, in the first form the same author name may appear under distinct names called synonyms, and in the second form distinct author names may have similar names referred to as homonyms (Shin *et al*., 2014). Author name ambiguity problem is closely related to other research fields like entity disambiguation (Bhattacharya & Getoor, 2007; Murnane *et al*., 2013; Chisholm & Hachey, 2015; Krzywicki *et al*., 2016; Oramas *et al*., 2016; Zhu *et al*., 2016), instance unification (Aswani *et al*., 2006), authority control Carrasco *et al*. (2016), Web appearance disambiguation (Bekkerman & McCallum, 2005), name disambiguation (On *et al*., 2005; Ferreira *et al*., 2012; Shin *et al*., 2014) object distinction (Zhu & Li, 2013), semantic matching (Giunchiglia & Shvaiko, 2003), record linkage (Christen, 2006; Kum *et al*., 2014), name variant problem (Maguire, 2016), name aliasing problem (Scholtes *et al*., 2016) and global names architecture (Pyle, 2016).

Generally, author name ambiguity is resolved by means of different publication attributes such as co-authors, title words, keywords, affiliations, references, abstract words, venues and publication years (On *et al*., 2005; Elliott, 2010; Ferreira *et al*., 2010; Esperidião *et al*., 2014). However, each DL is not providing all these attributes, they only provide scant information about these attributes and manual annotation is not possible at such a large scale. Furthermore, in recent years an ever increasing amount of publication data are being accepted and imported by DLs that makes names ambiguity problem more severe than in the past (Wang *et al*., 2011).

The problem of author name ambiguity is illustrated with the help of DBLP. Recently, when we search for an author name ‘C Chen’ in DBLP (a renowned DL), we get more than 20,000 different author names having same names or its variants, 137 different author names with exactly the same name ‘C Chen’. Methods that resolve name ambiguity problem in DLs are called author name disambiguation (AND).

We reviewed some of these techniques according to the defined inclusion/exclusion principal given in section 2.2 in this survey and their key characteristics are analyzed, such as the proposed methodology, which type of clustering or classification is used, applied similarity measures, metrics used for measuring the performance, uncertainty handled or not, used data set, capabilities of the techniques, evidence used and limitations of these techniques. We classified these AND techniques into five categories according to the applied AND technique. Details of these categories are given in section ‘A Classification for Author Name Disambiguation Techniques’.

1.1 Why need another AND survey?

Other reviews/surveys on AND can be found in the literature (Torvik & Smalheiser, 2009; Elliott, 2010; Ferreira *et al*., 2012). Torvik and Smalheiser (2009) refuted the need of a universal author name identifier or manual disambiguation for AND. They reviewed some AND techniques since 2007. Another description of techniques during 2004–2010 was done by Elliott (2010). She reassessed some individual efforts, some manual efforts and some projects that are working to resolve author name ambiguity problem like Authority, LCAF and The Names Project. A brief survey on AND was presented by Ferreira *et al*. (2012). They classified the AND techniques in to author assignment methods and author grouping methods. In author assignment method a reference is directly assigned to an author using some type of machine learning (ML)-based model, while in author grouping methods a similarity either predefined or proposed is used to group the citations. However, their proposed taxonomy is not comprehensive enough and is insufficient in the current literature. We propose a comprehensive taxonomy of AND techniques according to the used methodology—ML techniques (Wang *et al*., 2011; Huynh *et al*., 2013; Imran *et al*., 2013; Tran *et al*., 2014; Han *et al*., 2015; Seol *et al*., 2016) that have been further subdivided into three categories, and non-ML techniques (Fan *et al*., 2011; Wang *et al*., 2011; Tang *et al*., 2012; Shin *et al*., 2014) into two subcategories. Moreover, these all surveys reviewed the techniques up to the year 2012. There has been an extensive research effort in recent years on the solution of author name ambiguity problem. Figure 1 shows the increased trend of AND methods[8] within the last 12 years. So, there is a need for a survey that incorporates recent techniques.

[8] Indexed by Google Scholar on October 1st, 2016

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

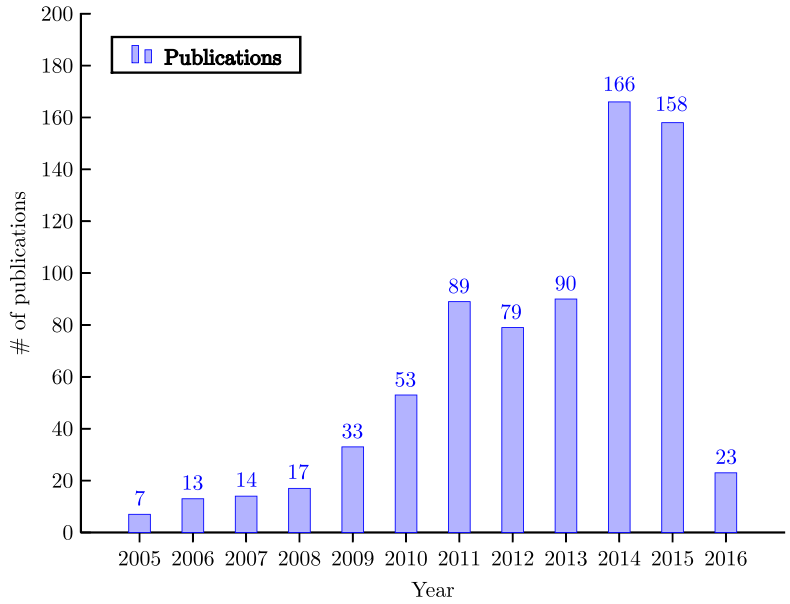


Figure 1 Number of author name disambiguation publications in last 12 years

A comparison of techniques from the past 6 years is done in the 'Selective Author Name Disambiguation Techniques'. Some open challenges and future research directions are given in 'Current Challenges and Future Research Directions'. Finally, we conclude the survey with a summary of presented techniques in 'Conclusions'.

2 Methods

To conduct this AND survey, we adopted the systematic literature review methodology proposed by Kofod-Petersen (2012). They devised a three-step methodology—planning, conducting and reporting.

2.1 Research questions

To achieve the aim of this review following research questions are thus posed as follows:

1. What are the existing techniques to AND?
2. What methodologies are followed in these AND techniques?
3. What implications will these findings have, when creating new AND systems?
4. What are the limitations of these AND techniques?

2.2 Research strategy

In search of the answers to the research questions identified in section 'Research Questions', relevant data (papers) for AND was collected using online DLs: ACM, DBLP, Google Scholar, IEEE Xplore, SpringerLink, CiteSeer and ISI Web of knowledge. The first keyword that we used for data retrieval is 'author name disambiguation'. Subsequently, we chose the keywords from these set of initial papers and found iteratively more papers. Keywords used in these searches were, for example, DL, name disambiguation, clustering, classification, synonyms, homonyms, citation analysis, bibliographic citation, common names, data integration, link discovery, mixed and split citations, and namesake resolution. Furthermore, we also used 'AND' and 'OR' of these keywords. We followed the literature search strategy of Moher *et al*. (2009). After this search strategy, we found 159 different papers. Then, we found 16 more papers by manually reading and noting the references of first set of papers and then retrieving those papers. We collected a set of 175 papers, out of which we selected only 31 papers that fulfills our designed selection criteria. So, finally 31 articles form the data set to answer the posed research questions in this

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

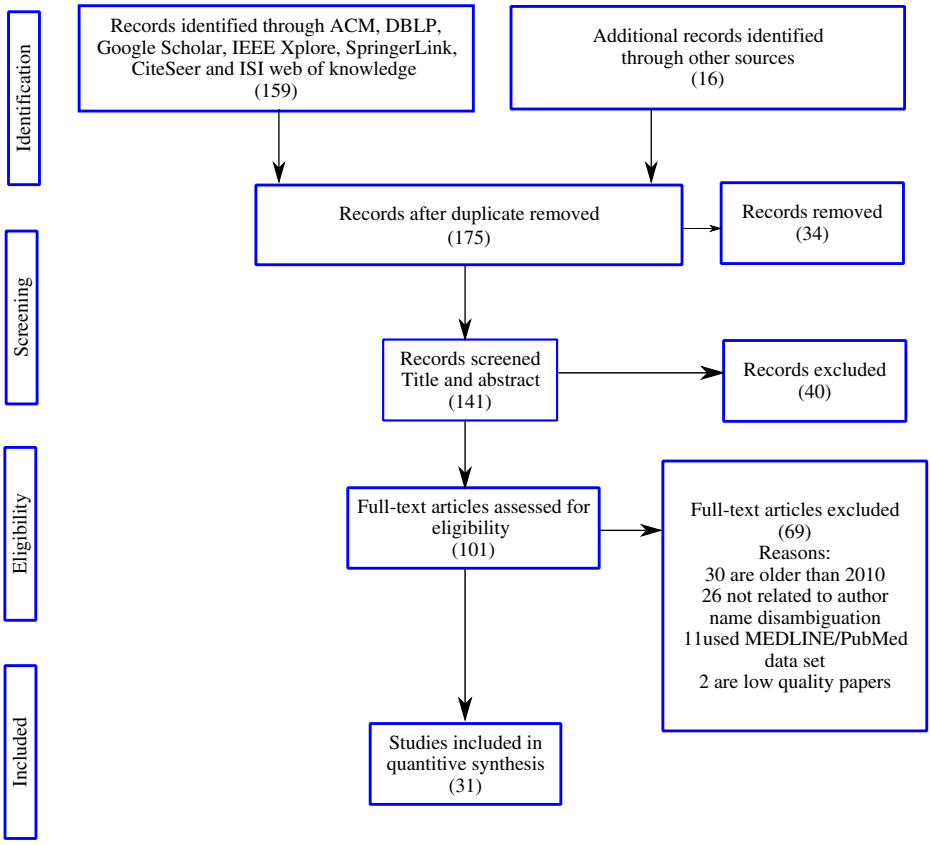


Figure 2 Literature review and selection process

survey. We strictly focused on recent works (past 6 years' only) that used author name ambiguity and we do not included works that used only MEDLINE/PubMed data sets and also not those that mainly focused on other research areas like record linkage, authority control, entity resolution and instance unification, etc., as our selection criteria. The details of every step can be shown in Figure 2.

2.3 Comparison criteria

We compared these AND techniques according to the following quality criteria. These quality criteria may not fully characterize AND techniques but with the help of these characteristics, we may be able to somehow compare the performance of several AND techniques. We briefly discuss one by one these quality criteria as follows.

a. Capability: AND techniques have two main problems: homonyms and synonyms. If a method handles both of these problems than it is better than the method that only deals with the single problem.
b. Evidence: It is related to the requirement of the attributes used in the method, either the primary evidence is required only or secondary evidence is also needed. The primary evidence means attributes readily available in citation records and secondary means auxiliary attributes that are not present in citation record and needs further processing.
c. Uncertainty: In AND, missing data about some attributes is called uncertainty. A technique is considered good if it is robust to these uncertainties.
d. Preliminary: This characteristic gives detail about the requirements of the human intervention like some techniques require user feedback, others require setting the threshold. Many techniques require Web data, some need complete data, others require total ordering of rules, and still others require citations should be greater than two. Techniques that require no preliminary data are good techniques.
e. No. of Ambiguous Authors 'K': Number of ambiguous authors 'K' is known in advance or not. Some techniques estimate the number of unknown authors. Those techniques are better which do not require

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

number of ambiguous authors in advance because in real world we have no clue about how many ambiguous authors in a DL.

f. Limitations: Limitations are some issues or drawbacks of the presented techniques.

2.4 Data sets and evaluation metrics

In AND three data sets—DBLP[9], Arnetminer[10] and BDBComp[11]—are used frequently and considered as bench mark data sets. They are publicly available and can be downloaded from respective websites. However, some researchers used other data sets such as Microsoft Academic Search, KISTI, ArXive, and MEDLINE. DBLP is the most used collection that is composed of 8442 citation records associated with 480 distinct authors belonging to 14 ambiguous groups, which means an average of ~18 citation records per author, as shown in Table 1. The original version of this collection was created by Han *et al*. (2004), and, with slight variations, it has been used in performance evaluation for various author disambiguation methods (Peng *et al*., 2012; Ferreira *et al*., 2014; Shin *et al*., 2014; Wu *et al*., 2014; Zhu *et al*., 2014; Han *et al*., 2015; Liu & Tang, 2015). Han *et al*. (2004) created this collection by collecting bibliographic citation records from DBLP and authors homepages. After that, they transformed author names to abbreviated forms consisting of the first name initial and the last name, and clustered the bibliographic citation records into ambiguous groups, each of which corresponds to the authors with the same abbreviated name.

An excerpt of the Arnetminer data set and its statistics are given in Table 2 and more details are found online. The original version of this collection was created by Wang *et al*. (2011). Then, Tang *et al*. (2012) manually checked, labeled and included more ambiguous authors to expand this data set. It is used with slight variations in many AND studies (Tang *et al*., 2012; Shin *et al*., 2014; Wu *et al*., 2014). Subsets of this data set have also been used in other works (Han *et al*., 2004, 2005, 2015; Ferreira *et al*., 2010).

BDBComp is relatively a small data set of 363 records belonging to 184 distinct authors, but it is very difficult to disambiguate as some authors have only one citation record. BDBComp data set statistics are given in Table 3 and this collection has also been frequently used in AND studies (Levin & Heuser, 2010; De Carvalho *et al*., 2011).

Metrics are important to judge the quality, performance, efficiency or progress of a process, product or plan. The number of Man of the matches or batting average of a cricket player is a well used metrics to rank cricket players, and surely for event sponsors to rate players. In all disciplines, metrics play pivotal role to access the performance. All Sciences including bibliometrics, scientometrics or informetrics use different quality metrics. However, for comparison we are using the following metrics to measure the performance of our proposed system with that of baseline methods.

Mostly, three metrics—K-metric, pairwise-F1 (PF1), and cluster-F1 (CF1) are used to measure the effectiveness of AND techniques (Pereira *et al*., 2009; Shin *et al*., 2014).

2.4.1 K-metric

K-metric is defined as the geometric mean of the average cluster purity (ACP) and the average author purity (AAP). ACP evaluates the purity of the algorithm-generated clusters with respect to the gold standard reference clusters, so it measures the amount of data misclassified into the wrong clusters by checking whether the generated clusters include only the publication records belonging to the reference clusters. AAP evaluates the level of splitting author information into several clusters, so it checks how fragmented the generated clusters are. ACP, AAP and K-metric are expressed in the following equation:

$$\mathrm{ACP}=\frac{1}{N}\sum_{r=1}^{R}\sum_{s=1}^{S}\frac{n_{rs}^2}{n_r},\ \mathrm{AAP}=\frac{1}{N}\sum_{s=1}^{S}\sum_{r=1}^{R}\frac{n_{rs}^2}{n_s},\ \mathrm{K}=\sqrt{\mathrm{ACP}\times\mathrm{AAP}} \tag{1}$$

here, N denotes the size of the citations in the collection, S the number of gold standard reference clusters manually generated, and R the number of clusters automatically generated by the Proposed Algorithm.

[9] http://dblp.uni-trier.de
[10] https://aminer.org/
[11] http://www.lbd.dcc.ufmg.br/bdbcomp

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Table 1 The DBLP data set

S No.	Name	No. of authors	No. of citation records
1	A. Gupta	26	577
2	A. Kumar	14	244
3	C. Chen	61	800
4	D. Johnson	15	368
5	J. Lee	100	1417
6	J. Martin	16	112
7	J. Robinson	12	171
8	J. Smith	31	927
9	K. Tanaka	10	280
10	M. Brown	13	153
11	M. Jones	13	259
12	M. Miller	12	412
13	S. Lee	86	1458
14	Y. Chen	71	1264
	Total	480	8442

Table 2 The Arnetminer data set

Names	Aut.	Rec.	Name	Aut.	Rec.	Name	Aut.	Rec.
Ajay Gupta	11	93	Lu Liu	17	58	Mark Davis	6	24
Bin Zhu	15	49	Cheng Chang	5	27	Michael Lang	4	17
Charles Smith	4	7	David Brown	25	61	Lei Chen	36	192
Michael Siege	6	54	David Cooper	7	18	Ning Zhang	31	125
David Wilson	5	67	R. Ramesh	9	46	Paul Wang	7	16
Eric Martin	5	85	Fei Su	4	37	Robert Allen	9	24
Yu Zhang	72	236	Gang Luo	9	47	S Huang	13	14
Sanjay Jain	4	216	Hui Fang	8	42	J. Guo	10	13
Hui Yu	21	32	Jie Tang	6	66	Ji Zhang	99	293
Xiaoyan Li	6	33	Yang Yu	19	71	Wen Jao	9	487
Jie Yu	9	32	John F. McDonald	2	34	X. Zhang	40	62
Bo Liu	65	306	Lei Wang	109	307	Yan Tang	6	27
Kai Zhang	28	82	Lei Fang	7	17	Wei Xu	47	153
Bin Li	65	306	Ping Zhou	18	36	Xiaoming Wang	14	41
M. Wagner	14	71	Rakesh Kumar	10	96	Lei Jin	6	16
Paul Brown	7	26	Shu Lin	2	76	Li Shen	6	65
Peter Phillips	3	13	Thomas D. Taylor	3	4			

Aut. denotes the number of distinct authors and Rec. represents citation records associated with that author.

Table 3 The BDBComp data set

S No.	Ambiguous group	Total records	Distinct authors
1	A. Oliveira	52	16
2	A. Silva	64	32
3	F. Silva	26	20
4	J. Oliveira	48	18
5	J. Silva	36	17
6	J. Souza	35	11
7	L. Silva	33	18
8	M. Silva	21	16
9	R. Santos	20	16
10	R. Silva	28	20

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Also, n_s is the number of elements in cluster s and n_{rs} the number of elements belonging to both cluster r and cluster s.

2.4.2 PF1

PF1 is defined as the harmonic mean of pairwise precision (PP) and pairwise recall (PR). PP is the fraction of publication records corresponding to the same author in the algorithm-generated clusters and PR is the fraction of publication records associated with the same author in the gold standard reference clusters. The PP, PR and PF1 measures are expressed in Equation (2), where $C(n, r)$ denotes the number of combinations of r elements from n elements: $C(n,r)=\frac{n!}{r!(n-r)!}, n \geq r$. Other parameters including r, s, n_s and n_{sr} are defined as before in Equation (1).

$$\mathrm{PP}=\frac{\sum_{r=1}^{R}\sum_{s=1}^{S} C(n_{rs},2)}{\sum_{r=1}^{R} C(n_r,2)}, \mathrm{PR}=\frac{\sum_{r=1}^{R}\sum_{s=1}^{S} C(n_{rs},2)}{\sum_{r=1}^{S} C(n_s,2)}, \mathrm{PF}-1=\frac{2\times \mathrm{PP}\times \mathrm{PR}}{\mathrm{PP}+\mathrm{PR}} \tag{2}$$

2.4.3 Cluster F1 (CF1)

CF1 is defined as the harmonic mean of cluster precision (CP) and cluster recall (CR), where CP is the fraction of the generated clusters that are equal to the reference clusters and CR is the fraction of correctly retrieved clusters from the reference clusters. The CP, CR and CF1 measures are given in the following equation:

$$\mathrm{CP}=\frac{R\cap S}{R}, \mathrm{CR}=\frac{R\cap S}{S}, \mathrm{CF}-1=\frac{2\times \mathrm{CP}\times \mathrm{CR}}{\mathrm{CP}+\mathrm{CR}} \tag{3}$$

2.5 Contributions

We reviewed these selected AND techniques, their prominent features, such as capabilities, limitations and briefly included some open challenges that needs researchers attention. We deliberately not include AND works that are older than 6 years now and have been discussed in some other surveys. We believe that this article will be useful for future researchers who are going to carry out research in AND domain. Moreover, our contribution complements the existing surveys by presenting: (a) an overall look of more recent AND techniques, (b) classification of AND techniques based on presented solution method, (c) complete synthesis of the AND techniques, and some open challenges and future research directions related to AND are also discussed.

3 A Classification of author name disambiguation techniques

AND techniques can be classified in many ways but we classified them into: ML-based techniques (Wang *et al*., 2011; Huynh *et al*., 2013; Imran *et al*., 2013; Tran *et al*., 2014; Han *et al*., 2015; Seol *et al*., 2016) that has been further divided into three subcategories, and non-ML-based techniques (Fan *et al*., 2011; Wang *et al*., 2011; Tang *et al*., 2012; Shin *et al*., 2014) into two subcategories. The proposed classification of selected AND techniques is shown in Figure 3.

ML techniques construct models based on prior observations which can then be used to predict the class of unseen data (Provost & Kohavi, 1998). The model is constructed using a learning process that mines valuable information about the data using the previous observations. ML methods usually receive a set of citations data for training and learns a model. The learned model is then applied to unseen data to attempt to guess the correct values. ML methods used in AND are of three types as described in the following paragraphs.

3.1 Supervised AND techniques

The first ML techniques are supervised techniques in which labeled training data are manually created and inputted to the classifier. The data consist of pairs of the form $\langle A_i, B_i \rangle$, where A_i is input feature vector and B_i is correctly labeled output class. The objective of learning function in supervised learning is to map input attributes to correct output class value. With the help of training data, a classification model is trained

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

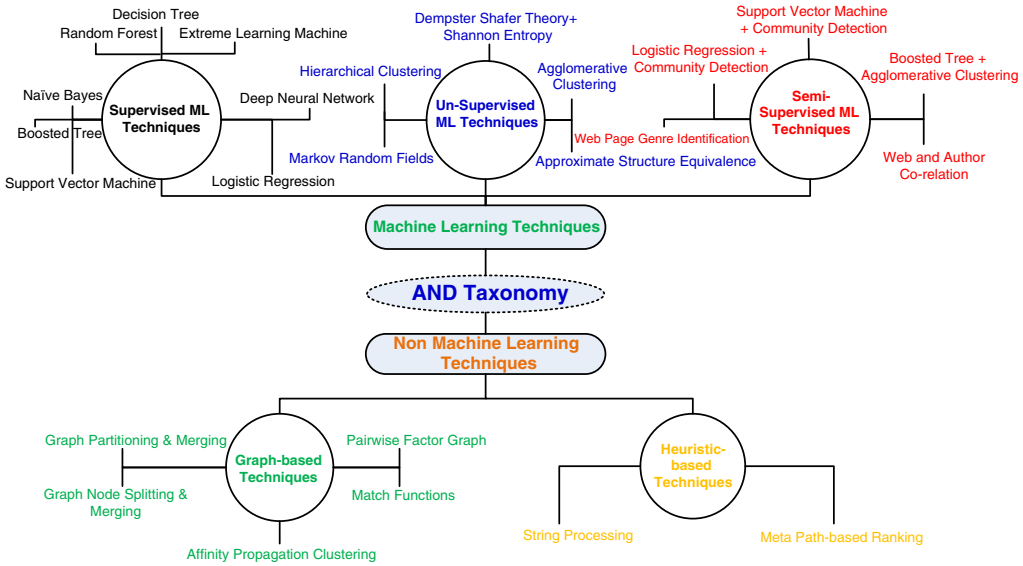


Figure 3 Proposed author name disambiguation taxonomy. ML = machine learning

and validated using some validation technique such as '*k*' fold cross-validation. Here '*k*' can be any number from 2 to *N*. Then the trained model is used to predict the output of unseen data. For example, the input can be a set of publication features like co-authors, title words, year of publications and venue information and the output can be the true author name class. True authors are target authors and false authors are homonym/synonyms authors of the target authors.

3.2 Unsupervised AND techniques

The second ML techniques are unsupervised techniques in which the learner is provided with input data, which has not been labeled. The aim of the learner is to find the intrinsic patterns in the data that can be used to determine the correct output value for new data instances. The assumption here is that there is a structure to the input space, such that certain patterns occur more often than others, and we want to see what generally happens and what does not. In statistics, this is called *density estimation*. A variety of unsupervised learning algorithms has been used in AND. In these techniques, some predefined similarity measures or some similarity functions are used for forming and finding the clusters of ambiguous author names. Key challenges to this type of technique are how to find the number of clusters and the value of the suitable threshold for similarity.

3.3 Semi-supervised AND techniques

The main disadvantage of the supervised technique is that they require a large amount of labeled training data. Similarly, finding number of hidden clusters and similarity threshold in unsupervised techniques is difficult. To overcome these drawbacks of supervised and unsupervised techniques, semi-supervised techniques are introduced. In these techniques, authors want to achieve good accuracy by using a small amount of labeled training data in conjunction with unlabeled data. In nearly all semi-supervised techniques one assumption made about the data consistency that data close to each other or having similar structure are likely to have the same label. Some good results have been reported in the literature for this type of techniques and in this regard, a variety of algorithms has been developed for AND in literature (Imran *et al.*, 2013; Zhao *et al.*, 2013; Maguire, 2016).

3.4 Graph-based AND techniques

These are only some techniques Tang and Walsh (2010) and Shin *et al.* (2014), which can be purely called graph-based techniques. Otherwise, the majority of techniques such as, Fan *et al.* (2011) and Wang *et al.* (2011)

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

form a graph and then find some similarity in them. Subsequently, they cluster these results by using ML techniques to disambiguate the author names. However, we have classified both these techniques under the umbrella of graph-based techniques. Some authors say that graphs are a natural representation of author name ambiguity problem and it takes into consideration the semantics of the problem. They represent author names as nodes and their co-author relationships as edges between them. After constructing graph some graph-based similarity measures are applied for disambiguation.

3.5 Heuristic-based AND techniques

When exact solutions are not possible or it is too slow to get the exact solution then heuristic-based techniques come into the scene and give solution quickly. It is worth noting here that these solutions may not be an optimal but they are near to an optimal. These methods are sometimes called short cuts, the rule of thumb or observation based. A heuristic function searches the solution space on the basis of available information and decides to go to the branch that approximates the exact solution. These methods also at the end may use ML techniques to resolve the ambiguity of author names.

4 Selective author name disambiguation techniques

In this section, selected AND techniques are elaborated, compared and their findings are presented.

4.1 Supervised AND techniques

Wang *et al*. (2012) presented a boosted tree classification method for name disambiguation that comprises of four steps. In the first step, name and affiliation filtering are done by matching first initial and last name and affiliation matches, whereas similarity scores for six different publication features are calculated in the second step. Then author name screening is done using false rate and in the last stage boosted tree classifier is applied to the manually crafted data set of 100 authors having 4253 citations in it. It cannot classify high false rate authors and requires manual checking. A deep neural network-based approach is proposed by Tran *et al*. (2014) to automatically learn features and disambiguate the author names from any data set. However, they tested their proposed solution with Vietnamese's data set that they prepared and used in their earlier study. The architecture of this solution has two main components. In the first component, data is taken as input and data representations are computed. These data representations can be prepared in many ways but authors' used string matching technique. According to their claim, these computations can be done on any data set automatically. The second component take the basic feature set as its input and learn the features in it is hidden layers to disambiguate the author name. The last layer of the feed forward deep neural network computes the probabilities to find whether two instances of the author name in a pair belongs to the same author name or not. They utilize the multi-column deep neural network technique to improve the generalization capabilities of the system that is similar to ensemble method bagging. Optimal no of hidden layers and no of units is a complex task and requires skill and experience.

Two extreme learning machine-based algorithms for AND are proposed by Han *et al*. (2015). First is one classifier for each name (OCEN) and the second is one classifier for all names (OCAN). In OCEN for each name a classifier is trained with the help of some attributes and when an unseen paper that is written by these ambiguous authors' is given to the classifier, it tells the identity of that author. They use a list of author names, title words of papers and venue title words as attributes for OCEN and extract features from these three attributes. Then, they reduce the dimensionality of features via principal component analysis. In the end, they formulate and solve the optimization problem using extreme learning machine method. In OCAN, they train a classifier that can predict whether the same author name in any given pair of entries in the bibliography referred to the same entity or two different entities. The idea behind this strategy is that an entity pair provides an abstraction from the concrete names. So, this classifier is not concerned with any specific name, which enables it to disambiguate all names. In this strategy, they use similarity of the two entities. They formulate the feature vector of similarities between author names, title words, and venue title words. They create enhanced feature extraction, which exploit additionally the information in the relationship of the entries referring to the same author name. Finally, they apply extreme learning machine

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

method on the proposed problem and find the solution. They compare its performance with support vector machine classifier and conclude that it has similar or better generalization performance. Missing data case is not handled in this strategy. If a single author or two homonyms share the same or similar title it would fail to distinguish.

A discrimination function that predicts true authors (target authors) and false authors (homonyms) using logistic regression on Web of Science data set is used by Onodera *et al*. (2011). They extract true author papers from 629,000 retrieved papers by using two stage filtering. In the first stage, they remove the retrieved papers if either their affiliation addresses has low similarity score to those of its source papers or there is no citation relationship between the venue of the retrieved papers and that of source papers. During the second stage, retrieved papers are manually judged and on the basis of this judgment a discrimination function of logistic regression is defined. The salient discrimination predictors are common co-author(s), the similarity of address, the similarity of title words, and citation relationships in venues between the retrieved and source papers. This method is not good for papers whose subject fields and affiliation addresses vary a little or do not vary at all. Five supervised ML techniques: Random Forest, Support Vector Machine, k-Nearest Neighbors, C4.5 (decision tree) and naive Bayes are proposed by Huynh *et al*. (2013) for solving Vietnamese author name ambiguity problem. They use Levenshtein similarity for calculating similarity between features. They propose a set of features from publications' data set that can be used to assist training classifiers. They test their model on three different data sets from the online DLs. It requires very specific neat and clean data set for the training of the models.

Table 4 shows the comparison of supervised AND techniques.

4.2 Unsupervised AND techniques

Wu *et al*. (2014) proposed algorithms for author name disambiguation that uses Dempster–Shafer theory (DST) in combination with Shannon entropy (SE). In the first step, some high-level features like affiliation, venue, contentisim, co-authors, citations, Web correlation and their co-relation similarities are calculated. In next step, these features are fused using DST and SE. On the basis of this information, belief and plausibility of each author is calculated. They get a matrix of pairwise correlation of papers. Each entry in this matrix is linked to a belief function and a plausibility function. Finally, they apply the DST-based hierarchical agglomerative clustering algorithm for author name disambiguation. In the process of clustering they use three different convergence conditions for clustering algorithm namely: pre-set number of clusters, the number of available evidence and distance between clusters.

Cognitive maps of psychology and structural equivalence of network analysis-based knowledge homogeneity scores are used to recognize the ambiguous author name in bibliographic, particularly common surnames in China, is presented by Tang and Walsh (2010). Their basic assumption is that every author has its particular set of the knowledge base at a particular time. During that time, two authors having the same knowledge base are considered the same author whereas two authors having different knowledge base are considered two distinct authors. Finding true structural equivalence in the real world is rare. Hence, approximate structural equivalent (ASE) is used instead of equivalence, such that authors within a structurally equivalent cluster are more similar to each other than to those outside of that cluster. If these structurally equivalent records contain author names with the same (or similar) family name and first initial, these similar authors are taken to be the same authors. To find the ASE, a knowledge homogeneity similarity (KHS) score is calculated that is based on the summation of shared references, the forward citations of each reported reference, and the minimum number of references reported by the two articles. Rarer references are given higher weights. Once the KHS matrix is constructed, hierarchical clustering with single linkage is done to distinguish the groups. These methods do not perform well when references of a citation are not available.

A unified probabilistic framework to address the name disambiguation problem specifically homonyms in DLs is proposed by Tang *et al*. (2012). They formalize the disambiguation problem using Markov random fields, in which the data are cohesive on both local attributes and relationships. In this approach, they suggest an algorithms that automatically estimate the number of unknown ambiguous authors and the required unknown parameters. The basic idea behind this work is that a similar content and similar

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Table 4 Supervised author name disambiguation techniques comparison

Ref.	Similarity	Methodology	Performance metric	Uncer.	Data set	Evidence	Capability	Limitations
Wang *et al*. (2012)	Cosine, eigen decomposition	Boosted tree classification	Precision, recall and misclassification error	No	Not standard	Author, venues, keyword, title word, abstract, subject category	Homonyms	How to decide the splitting point and how to control the size of the tree
Han *et al*. (2015)	TF-IDF	ELM-based classification	Mean and SD of test accuracy	No	Self-designed DBLP	Authors, co-authors, title words, venue	Homonyms	It is difficult to decide network structure and its parameters
Tran *et al*. (2014)	Jaccard, Jaro, Levenshtein, Smith-Waterman, Jaro-Winkler, Mogne-Elkan	Deep neural network	Accuracy, error rate	Yes	IEEE Explore Vietnamese authors, MAS, and ACM	Author name, affiliation, co-author, interest keywords	Both	It requires retraining the model if some parameters are changed. Many models for each author
Onodera *et al*. (2011)	Self-defined	Logistic regression	Precision, recall, accuracy	No	WOS	Co-authors, affiliation, citation relationship, title words, co-citations, year of publication, country	Homonyms	Cannot solve transitivity problem. Fails if subject fields and affiliation do not vary too much
Huynh *et al*. (2013)	Jaccard	Random forest, SVM, K-NN, decision trees, Bayes	Accuracy	No	ACM Vietnamese DS, IEEE, MAS	Authors, co-author names, authors affiliations, keyword similarity	Both	Cannot handle outliers

From left to right: Ref. is the references of the paper, Similarity used, what Methodology followed, Performance Metrics, Uncer. refers to either Uncertainty handled or not, Data set used, Evidence used, Capability of the method and Limitations.

relationship belongs to the same author. They claim that their approach can achieve better performance in name disambiguation than baseline methods because the approach takes advantage of interdependencies between paper assignments. In the first step, they assign six attributes (Title word, venue name, Publications' year, Abstract, Authors, References) to each paper that they acquire from the online DLs. They define five types of relationships namely, co-pub venue, co-author, co-citation, constraints and time of co-authorship among these papers, and gave unknown weights to these relationships which they find in the later stages. They transform the content-based information and structure-based information into the hidden Markov random fields (HMRF) as feature functions. They define an objective function as the maximum a posteriori configuration of the HMRF. They use Bayesian information criterion to estimate the true number of authors. To find the unknown parameters they devised an algorithm that first initializes and assigns random values to unknown parameters and then updates these values for each function values to optimize the objective function.

Another algorithm that not only disambiguates author name problem but it also reconstructs the h-index of the authors is proposed by Schulz *et al*. (2014). They apply this algorithm on a large scale data set of Web of Science. This algorithm consists of two main steps. In the first step, they calculate the pairwise similarity between all papers on the basis of number of shared co-authors, self-citations, common references and the number of papers citing both publications. Calculating these pairwise similarities between papers, they construct a link between those publications that have similarity score greater than some threshold. Their aim is to find a distinct author that they find by knowing each connected component. These connect links form a number of clusters and a new similarity between these clusters is calculated. A link between these clusters is constructed if this similarity is above a threshold as in the previous step. Then these connected clusters are merged. These clusters are the set of papers by a unique author. They optimize and validate the disambiguated authors by re-constructing their h-index. Information about h-index is necessary for this system, but new researchers have no h-index information and this system does not disambiguate them well.

A technique named 'Fast Multiple Clustering' is presented by Liu *et al*. (2015) in which three-step clustering is used to disambiguate author names. In the first step, co-authors are used to finding clusters of the authors in which multiple relations such as paper related to authors and papers related to other papers are found. On the basis of these relations, the related papers are clustered in different clusters. Further, bigger clusters are formed on the basis of similarity between title words. Finally, venue information is used to cluster the papers of the authors who tend to publish in the same venue but may have different title words. Authors claim that this technique is better in PF1 and faster as compared to three existing techniques. Another advantage of their proposed technique is that it can be used for almost all bibliographic databases.

Comparison on different quality criteria of unsupervised AND techniques is given in Table 5.

4.3 Semi-supervised AND techniques

A hybrid name disambiguation framework that not only used the traditional information, co-authors, but also Web page genre information is proposed by Zhu and Li (2013). This framework consist of two main steps: Web page genre identification and re-clustering model. In Web page genre identification, returned pages are classified either home page of an author or not. Those records that found at the authors' home page belonged to him and are disambiguated. The remaining ambiguous records are disambiguated using co-authors' information. Some records that neither disambiguated from co-authors' names nor from Web information are sent to re-clustering model. Then they build a graph 'G' using all citation records in which each vertex represents a citation record and each edge denotes the same co-authors relationship or they are of the same domain. If there are many links present between two vertices, then they are considered to be related to each other. They transfer this graph into a similarity matrix by using multi-dimensional scaling algorithm. This algorithm detect similarities among objects. They construct two-dimensional matrices for co-authorship and topic-relationship and calculate the distance between two vertices with the help of Euclidean distance. If the distance between citations is less than a specific threshold then they are considered by the same author. One author might have more than one citation record on one personal page and

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Table 5 Unsupervised author name disambiguation techniques comparison

References	Clustering	Similarity	Methodology	Performance metric	Uncer.	Data set	Evidence	Capability	Limitations
Wu *et al.* (2014)	Hierarchical agglomerative	JC, LD, Cosine	Dempster–Shafer Theory with Shanon Entropy	K, Pairwise-F1	Yes	DBLP	Co-author, Affiliations, Venues, Content Similarity, Citations, Web co-relations	Both	Low accuracy
Schulz *et al.* (2014)	2-Step agglomerative	Own defined	2-step agglomerative clustering	Precision, h-index error	Yes	WoS	Author, shared citations, self-citations	Both	H-index is not known for new authors
Tang and Walsh (2010)	Hierarchical agglomerative	Knowledge homogeneity score (KHS), Self-defined similarity	Approximate Structure equivalence-based KHS	Accuracy	No	Limited (own designed)	Author name, references	Both	Homogeneity threshold varies from subject to subject
Tang *et al.* (2012)	–	Cosine	Unified Probabilistic framework with Markov random fields	Pairwise accuracy, precision, F1	No	Self-designed	Authors, title words, abstracts, reference, year, venue	Homonym	If relationship information is not available then this method would not perform well
De Carvalho *et al.* (2011)	Hierarchical	Fragment comparison, cosine, TF-IDF	Heuristic based	K metric	Yes	BDBComp, synthetic	Authors, title words, venues	Both	It fails if similar author name but does not have similarity, not have co-authors and not match with any existing cluster
Liu *et al.* (2015)	Hierarchical	Cosine	Hierarchical clusters of relations between authors and papers	PP, PR, PF1	No	DBLP	Authors, Title words, Venues	Both	It fails if authors with the same name have similar research fields

From left to right: Reference of the paper, Type of clustering, Similarity used, what Methodology followed, Performance metrics, Uncer. refers to either Uncertainty handled or not, Data Set used, Evidence used, Capability of the method and Limitations of the methods.

have other different citation records on another personal page that represent as two distinct authors which cannot be fully handled by this approach.

Levin *et al*. (2012) presented a semi-supervised two-stage method to disambiguate authors in DLs. In the first stage, they used citation-based rules along with some other simple rules to create labeled training data automatically. Then these initial clusters are being used as a bootstrap to learn the similarity metrics for the second stage of agglomerative clustering. They used a large number of old as well as new features for measuring the similarity between pairs of publications and solve the optimization problem to find disambiguated authors. They evaluate this model on a data set collected from the Thomson Reuters Web of Knowledge. Ferreira *et al*. (2014) proposed an enhancement to their previously proposed technique in 2010 named SAND (self-training associative name disambiguation) that consists of three steps. Pure but fragmented clusters are generated in the first step using co-authors heuristic. Representative clusters from the first step are used for the training of the model. In the last step, finding associative name disambiguator and training of the model that detects the appearance of the new authors. Remaining clusters that are not used for training of the model are used for testing purpose. They test SAND on DBLP and BDBComp data sets and achieve better results than two supervised and two unsupervised approaches. Finding the pure clusters in the first phase is a tough task.

A semi-supervised approach for AND that utilizes Microsoft academic search data are proposed by Zhao *et al*. (2013). They pre-processed data set and found many useful features. They construct a co-author-based bibliographic network and apply community detection algorithm. Support vector machine and some other ML models are used to handle uncertainty in the data. They boost the performance of these models with the help of a self-taught procedures. Lastly, they merge the results from different models and achieved 0.98717 mean *F*-score on Microsoft academic search data set provided by KDD cup 2013. Very recently, an ethnicity sensitive method that mainly comprises of three parts is presented by Maguire (2016). In the first part phonetic-based blocking for similar author signatures is done. After blocking, supervised ML-based linkage function is proposed that exploited the ethnicity sensitive information. Author names are divided into seven groups of ethnicity as White, Black, American Indian or Alaska Native, Chinese, Japanese, Asian or Pacific Islanders and others using. For a pair of given names the probability of the ethnic group for both the names is estimated and on the basis of this information the origin of the author is predicted. Finally, hierarchical agglomerative clustering is done on the basis of a distance between two pairs of publications' linkage function. Web co-relations' and author co-relations'-based approach to measure similarities between publications for AND is presented by Peng *et al*. (2012). They use two assumptions, citations on the same Web page are related to the same author and citations with same rarer authors belong to the same author. They measure these both type of co-relations by pairwise similarity metrics by using modified sigmoid function, cosine metric and name popularity metric.

Gurney *et al*. (2012) developed an algorithm that address the issues of discarded records due to null data fields and their resultant effects on recall, precision, and F-measure. They implement a dynamic approach to similarity calculations based on all available data fields. They use the Tani Moto coefficient for similarity calculations of (a) title words (b) abstract words (c) last names and first initials of co-authors (d) cited references in whole-string form (e) normalized author keywords (f) normalized indexed keywords (g) normalized research addresses (h) venue. They also include average author contributions and age difference between publications, both of which has meaningful effects on overall similarity measurements, resulting in significantly higher recall and precision of returned records. Logistic regression is then applied to find the unknown weights to all the parameters. In the final stage, authors are found with the help of Blondel community detection algorithm. This algorithm take into account the weighted edges of a network and assign each node to a specific community based on the surrounding nodes and its edge weights. Logistic regression predicts the probability whether two publications are from the same author or not, but the community detection algorithm works on the entire interconnected network of nodes or publications and identifies the communities of papers belonging to unique authors. The results are presented from a test data set of heterogeneous catalysis publications and demonstrate significantly high average F1 scores and substantial improvements compared with previous stand-alone techniques. Null combination code (NC) is the presence or absence of shared meta data in a citation. It is a numeric string where each number refers to a field present in the citation. For example, NC code 124 represents that title, abstract and author assigned

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

keywords are present. A zero in this code means respective field is not present in the citation. This method requires pre-checking of records and calculation of *NC* code manually that is a tedious job.

A solution for AND is proposed by Imran *et al*. (2013) that can also be used as a wrapper service for DLs. They use selected features of an author and citation record to disambiguate author names using unsupervised techniques. First of all, they collect citation records of an ambiguous author name in which both the mixed citations and split citations exist. They make clusters of their citations on the basis of their disciplines and then divide these clusters into smaller ones using co-authors information. Afterward, these small clusters are fused together on the basis of remaining evidence. In this phase, two clusters are fused only if their distance is reduced to a certain threshold limit. Meanwhile, they interact with users to further purify the retrieved clusters, so users' feedback directly influences the results of the disambiguation. Wang *et al*. (2014) proposed a unified semi-supervised framework which is capable of solving both homonyms and synonyms. This framework uses semi-supervised approach to the solution of author names where there are no training data are provided. Multi-aspect similarity indicator and a support vector machine are applied to fuse the attributes. In the final step, a self-taught method is presented to resolve the ambiguities in the authors to enhance the performance of the disambiguation system. A comparative analysis of semi-supervised AND techniques is done in Table 6.

4.4 Graph-based AND techniques

Two multi-level scalable algorithms for AND are proposed by On *et al*. (2012). First is a multi-level graph partitioning (MGP) algorithm, and the second is a multi-level graph partitioning and merging (MGPM) algorithm. Input data are represented in the form of affinity graph, where each entity denote the node and relationship between two nodes represented the edge. Edge weights between two entities are calculated using the term frequency inverse document frequency of all word tokens. These graphs are then level by level approximately partitioned into the smaller graphs using MGP schemes. Three steps are followed for the MGP algorithm: scaling-down, partitioning and scaling-up. During the scaling-down, the size of the graph is repeatedly decreased; in the clustering, the smallest graph is partitioned and in the scaling-up, partitioning is successively refined to the larger graph. MGP has somewhat low accuracy as compared to MGPM. To overcome this drawback they propose MGPM algorithm that is faster but more accurate graph partitioning method. The MGPM algorithm is also a MGP algorithm but it allows multi-resolutions at variant levels. After every stage, the algorithm decide to go next level only if the partitioning step produces more inter-cluster edges than intra-cluster edges. Otherwise, the MGPM algorithm stop at the current level. In the end, the MGPM algorithm generate different levels of each branch and various resolutions for each leaf node. After dividing graph into smaller sub-graphs, they combine the sub-graphs together if two sub-graphs has the biggest normalized cuts. They repeatedly merge the sub-graphs until the number of sub-graphs is equal to the given number of clusters '*K*'. The drawback of this method is that it assumed predefined no of clusters.

Collaboration network of authors along with syntactic similarity between author names to disambiguate author names in three different subsets of DL data sets is presented by Levin and Heuser (2010). They assume that two syntactically similar authors are same if there are a close relationship and small distance between them. They make a graph by considering two type of nodes; author nodes and paper nodes, edges are relationships between papers and their writers. They define multiple metrics to measure the closeness and distance between authors. Finally, they define five matching functions based on these metricsone for syntactic matching and fours for relationship matching. Their research results depict that this approach significantly improve the performance of just syntax-based similarity measures. A graph-based method called graphical framework for name disambiguation is proposed by Fan *et al*. (2011). In it they model the relationships among publications using undirected graphs. Authors are represented with a vertex and an edge shows a coauthor's relationship in this graph. Then, they solve homonyms problem by iteratively finding valid paths, computing similarities, clustering with the help of affinity propagation algorithm and in the last step using user feedback as a complementary tool to enhance the performance. They simulate using PubMed and DBLP data sets, and results reveal that their proposed approach is better both in

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Table 6 Semi-supervised author name disambiguation techniques comparison

Ref.	Similarity	Methodology	Performance metric	Uncer.	Data set	Evidence	Capab.	Limitations
Zhu *et al*. (2014)	Multi-dimensional scaling	Web page genre identification based graph re-clustering	F1	Robust	DBLP	Explicit web genre	Both	Slow as using web information
Gurney *et al*. (2012)	Tani Moto coefficient	Logistic regression & community detection algorithm	Precision, recall, *F*-measure	Yes	Self-designed WoS	Title words, abstracts, keywords, citations, difference in years of publication and average author contribution	Homonyms	Requires too many parameters
Imran *et al*. (2013)	Levenshetien	Heuristic-based, unsupervised and adaptive	Precision, recall, F-1	No	DBLP	Authors, affiliations, title words, venues, home page	Both	It involves user on multiple stages
Zhao *et al*. (2013)	Cosine, TF-IDF, LDA, tanimoto	Community detection and SVM algorithms	Precision, recall, F-1	Yes	MAS	Title words, venue, keywords, affiliations, publication year	Both	Complex rules and topic modeling needed
Maguire (2016)	Cosine, Jaro-Winkler, TF-IDF	First gradient boosted tree applied on similar authors and then agglomerative clustering is performed	Pairwise precision, recall, F-1	No	INSPIRE	Authors, affiliations, title words, venues, abstract, keywords, collaborations, references, subject, year difference	Both	Not applicable to all DL's data set
Peng *et al*. (2012)	Modified sigmoid function, cosine, name popularity measure	Web and authorship correlations	Pairwise precision, recall, F1	No	DBLP	Authors, title words, venues	Both	Inherently slow as using web info
Levin *et al*. (2012)	TF-IDF, cosine	Self-citation clustering rules with other rules	Pairwise precision, recall, F-1	No	WoK Thomson Reuters	Title word, venue, authors, addresses, affiliations, subject categories, citations languages, year, combinations of these	Both	Too much processing for large set of features
Ferreira *et al*. (2014)	Cosine, Euclidean	First pure clusters of data are found and model is trained, tested on them	K, pairwise-F1	No	DBLP, BDBComp, SyGAR	Authors, title words, venues	Both	Threshold selection is a big issue
Wang *et al*. (2014)	Cosine, Tanimoto, TF-IDF	Similar authors share co-authors and have high topic similarity	F1	No	MAS	Title words, affiliation, keyword, venue	Both	Fails in case of sole authors

From left to right: Ref. is the references of the paper, Similarity used, what Methodology followed, Performance metrics, Uncer. represents either Uncertainty handled or not, Data set used, Evidence used, Capab. refers to the Capability of the method and Limitations.
DL = digital libraries.

precision and recall than DISTINCT—a baseline approach. This approach does not handle outliers and fail in the case of sole authors.

In a study by Liu and Tang (2015), a problem and knowledge domains-based bi-relational network (BRN) framework for AND is proposed. The problem domain is used to construct basic BRN and knowledge domain is exploited to elaborate the covert aspects of the network. Subsequently, they use a random walk with a restart to find the closeness between BRN nodes and affinity propagation algorithm for clustering the obtained results from the previous step. A graph-based method—ADANA is proposed by Wang *et al*. (2011), in which they modeled pairwise factor graph that can be used to integrate several types of features as well as user feedbacks into a unified model. They define three types of feature functions: document pair, correlation and constraint-based feature functions. In document pair feature functions, they found known relationships from publications. In correlation feature functions, they found some hidden features with the help of known functions and in the constraint-based feature functions, the user is involved in finding the unknown features. Finally, they exploit active selection of the users' corrections in an interactive mode to improve the disambiguation performance after some preliminary clustering results. They exploit some additional information such as affiliation, references, etc. that is not present in every DL.

A topological–collaborative approach is presented to solve homonyms in DLs by Amancio *et al*. (2015) in which they used topological features of the collaborative graph along with weighted collaboration patterns among authors. In this technique, the first step is the formation of the weighted network among authors according to the strength of their collaborations and then characterized the network using several topological features such as neighborhood degree, neighborhood strength, clustering coefficient, average shortest path length, hierarchical measurements and locality index. Fuzzy *K*-NN is used in the classifier to disambiguate different homonyms. Authors validated proposed technique on a data set extracted from the arXiv repository and concluded that topological features enhance the accuracy of their results. Among these topological features, the average shortest path length is the most prominent feature for disambiguation. A graph framework for author name disambiguation (GFAD) is presented by Shin *et al*. (2014) that exploit authors, co-authors and paper title words information. They claim that this framework is robust and domain independent because it requires only author names, co-author names and paper titles information that is surely available in all DLs. They model the author name and co-author names in an undirected graph, where a vertex denoted an author name and an edge indicated a co-author relationship. Homonyms are resolved by splitting the vertex that is common to multiple non-overlapping cycles so that each cycle corresponded to a unique author. The heteronymous name problem is handled by merging multiple author vertices into one by identifying those vertices that actually represent a single author with different names. Moreover, unlike other related studies, outlier issues are also handled in GFAD. It remove outliers by comparing similarity among outliers with the help of cosine similarity. Experiments are conducted on two real world data sets of DBLP and Arnetminer. This method fails when there is less or no title word similarity in two papers written by sole authors and also fails in the case of very ambiguous author names. Furthermore, as it used Johnson (1975) algorithm for cycle finding, so it is computationally too expensive. All these graph-based methods in this section are compared in Table 7.

4.5 Heuristic-based AND techniques

INDi a solution for the existing cleaned DLs is presented by De Carvalho *et al*. (2011). Their proposed solution utilize similarity among bibliographic records and group the new records to authors with similar citation records in the DL or to new authors when the similarity evidence is not strong enough. Some particular heuristics are used for checking whether references of new citation records belong to pre-existing authors of the DL or if they belong to new ones (i.e., authors without citation records in the DL), avoiding running the disambiguation process in the entire DL. They run simulations on BDBComp collection and synthetic DSs to assess the effectiveness of their method. They compare its performance with an unsupervised method. They use synthetic data for incremental disambiguation purpose that generated data from existing data and does not portray the real world situation where many new authors also publish.

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Table 7 Graph-based author name disambiguation techniques comparison

Ref.	Graph	Similarity	Methodology	Performance metric	DS	Evidence	Capab.	Limitations
On *et al*. (2012)	Co-authorship	TF-IDF	Graph partitioning and merging	Precision, recall, F1	4 Small to large DS	Documents	Both	Number of Clusters Required
Shin *et al*. (2014)	Co-authorship	LCS, Cosine	Graph Node Splitting & Merging	K-metric, Pairwise-F1, Cluster-F1	DBLP & arnetminer	Co-authors, title words	Both	Unable to disambiguate sole authors
Fan *et al*. (2011)	Co-authorship	User defined	Affinity propagation clustering	Pairwise accuracy, precision F1	Small standard DS	Co-author graphs, user feedback	Homonyms	Slow due to user feedback
Levin and Heuser (2010)	Author social graph	Levenshtein, trigram	Syntactic relationship match function	Accuracy, precision, F1	Cora, BDBComp, DBLP	Authors, title words	Both	If no relationship or low syntactic similarity then fails
Wang *et al*. (2011)	Pairwise factor graph (PFG)	TF-IDF, cosine	PFG and interaction of user to enhance the performance	Precision, recall, F-measure	Publication, web page and news stories	Citations, Co-authors, co-venues, co-affiliations, co-aff-occur, title word-sim, co-homepage	Homonyms	Not scalable due to huge number of path calculations
Liu and Tang (2015)	Bi-relational	Graph-based closeness	Bi-relational network is created, closeness between nodes is Found for clustering	–	Citations	Authors, title words, venues	Both	Knowledge base is required that is not available in some DL

From left to right: Ref. is the references of the paper, Similarity used, what Methodology followed, Performance metrics, Data set (DS) used, Evidence used, Capab. refers to the Capability of the method and method Limitations.
DL = digital libraries.

Table 8 Heuristic-based author name disambiguation techniques comparison

Ref.	Similarity	Heuristic	Data set	Limitations
Chin *et al*. (2014)	String comparison	String processing	MAS	It only finds duplicates. Not tell the No. of ambiguous authors
Liu and Tang (2015)	Levenshtein, Soundex	Meta path-based ranking	MAS	It fails on sole authors cases
De Carvalho *et al*. (2011)	String Comparison	Publication features	BDBComp	No real world data set is used
Santana *et al*. (2015)	Own defined	Publication features	DBLP, BDBComp, KISTI	Finding optimal threshold is a difficult task

From left to right: Ref. is the Reference, Similarity used, what Heuristic followed, Data set used and Limitations.

A name matching framework for author name disambiguation for Microsoft Academic Search data set is proposed by Chin *et al*. (2014) and it realize two implementations. This method has six stages in total. In the first stage, they decide that either the name is a Chinese name or non-Chinese. They group Korean, Singaporean and Vietnamese names into Chinese. They built two dictionaries of Chinese names. In the second stage, they clean and preprocess the citations data. In the third stage, they make blocks of similar author names. Author name blocks are created using a dictionary of (key, value) pairs. In this blocking strategy, a name is used as a key and the number of its occurrence as a value. For instance, an author name 'Ajay Kumar Gupta' has following keys. 'Ajay', 'Kumar', 'Gupta', 'Ajay Kumar', 'Ajay Gupta', 'Kumar Gupta' and 'Ajay Kumar Gupta'. In this blocking scheme, the order of the names does not matter. During this stage, they take care of low recall and high false positives that both might degrade the performance of the overall AND system at later stages. In the fourth stage, they identify duplicates in the blocks. After identification, they split some names that are wrongly included in the blocking stage. In the fifth stage, they link the names of authors to their identifiers. Lastly, they merge the results of two predictions by using background information and a filtering process. This merging step boosts the F1 score.

Santana *et al*. (2015) proposed a combination of the domain-specific heuristics to disambiguate authors in bibliographic databases. The authors assume that a researcher's publication profile may be characterized by the terms of one's citations. The co-authors form the collaborations networks and title and venue terms portray one's research interests. Using these three domain-specific heuristics a similarity function is proposed and if the similarity between two compared authors is greater than a specified threshold than these authors are included in the training data set. It is also assumed that if there is a low similarity between a citation and its most similar group then this is a new author which is not already present in the training data set. In the final step optimal threshold for different parameters are estimated using the standard procedure of cross-validation. Ranking-based name matching algorithm and a system called RankMatch is proposed by Liu and Tang (2015). RankMatch has four steps. The first step is preprocessing in which some data cleansing activities are done like noisy first and last name and mistakenly separated or merged names are identified. The next stage is *r*-step in which recall of the system is enhanced. The next stage is the *p*-step that improves the precision of the system. In this step, they calculate different meta path-based similarity measures. Then these ranking-based measures combine with string-based measures are used to remove the conflictory names from the results of the previous stage. In the last, post-processing stage, the unconfident names are removed from the system despite the fact that names are compatible and they might have acceptable meta path similarity between names. They achieved good F1 score on MAS data set. Comparison on different quality criteria of heuristic AND techniques is given in Table 8.

5 Critical discussion

Supervised AND techniques require representative labeled training data. Creating labeled training data is resource and time consuming and accurate labeling is often hard to achieve. For example, obtaining sufficient training data for AND in bibliography might require weeks to months to collect and then label that data. In some cases even in manual labeling, the data may not be correct due to lack of bibliographic evidence. On the other hand, According to Ferreira *et al*. the performance of supervised AND techniques is

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

relatively better than unsupervised techniques Ferreira *et al*. (2014). In a supervised study by Wang *et al*. (2012), author affiliation history is required that is usually not available in DLs. In Han *et al*. (2015), their proposed technique require a large number of training examples to train their model that is not available for a new or a rising researcher. Similarly, Tran *et al*. (2014) present a technique that requires retraining of the model if some parameters are changed or new citation have been inserted in the DL.

Unsupervised AND techniques do not require any training data set. However, unsupervised learning algorithms also have some drawbacks, some of them are as under: (a) it is difficult to decide which type of clustering best suits to the problem at hand, (b) no of clusters (K) is not known in advance, (c) when to stop the clustering process is not known and (d) these techniques are not as efficient as supervised ones.

The technique proposed by Tang and Walsh (2010) used homogeneity score for clustering and its threshold vary from subject to subject which is very difficult to find. Tang *et al*. (2012) technique performs poorly when relationship information is not available.

Semi-supervised AND techniques also require some training data as well as a number of clusters in advance. However, semi-supervised techniques like Zhu *et al*. (2014) and Peng *et al*. (2012) require Web information that slows down their progress in DLs. Whereas, Imran *et al*. (2013) proposed method need users feedback for higher accuracy of the proposed system.

Graph-based AND techniques require fewer attributes than other techniques. These techniques has a rigorous mathematical foundation and considered the semantics of the citations for disambiguation. GFAD is unable to disambiguate very ambiguous author cases and authors having no common authors among them.

Heuristic-based AND techniques are sometimes very efficient and give a solution in a very straight forward manner. However, it is possible these techniques might sometimes not give an accurate solution. These techniques do not have a rigorous mathematical foundation and consider some previous heuristic like coauthor inclusion, string processing or meta path-based similarity in AND. Liu and Tang (2015) proposed method fail in the case of sole authors.

6 Current challenges and future research directions

In spite of the development of a variety of AND techniques, a little effort has been focused on some of the challenges of the author name ambiguity problem. These challenges are the main reason behind the unreliable AND solutions. We discuss some of them in following paragraphs.

Sole author cases: Majority AND techniques considered that ambiguous author name can be disambiguated by knowing his or her co-authors identity. However, these studies failed or produced inferior results when there were citations that only written by the sole author. This case even became more worst when these sole authors had published only one or a few papers (Han *et al*., 2004; On *et al*., 2006; Levin & Heuser, 2010; De Carvalho *et al*., 2011; Fan *et al*., 2011; Shin *et al*., 2014). Some new features or techniques need to be explored to handle these types of authors such as reference similarity, affiliation information or some hidden topics of the publications.

Imbalance distribution of publications: According to a study by Torvik and Smalheiser (2009), 46% of the authors have written only one publication in their entire career. The authors' distributions of publications follow power law distribution. According to this law, many authors have only a few publications and a few authors have many publications. This imbalance citation distribution more hardens the AND problem.

Scalability: The majority of existing AND techniques are only applied to specific and limited scale data set. When they are applied on a larger scale either they loose performance or even fail to disambiguate. Due to the rapid increase in publications by academics in recent years, scalability of the AND techniques is also a challenging issue that needs attention and resolution. Schloss Dagstuhl Leibniz-Zentrum fr Informatik GmbH announced a project titled 'Scalable Author Disambiguation for Bibliographic Databases 2015–2018'[12] in which scalability is the main focus of their project. In this regard new AND systems should be validated on huge data sets.

[12] https://www.dagstuhl.de/ueber-dagstuhl/projekte/autoren-disambiguierung

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

No standard representation of names: Presently, every publisher has his own set of rules for author name representation which makes AND more challenging and difficult to resolve. Some writes as 'FirstName, LastName', others write as 'LastName, FirstName', still others write as 'FirstNameInitial, LastName'. To make some cases of name ambiguity somewhat simpler and easier publishers should come to an agreement on the standardization of representing names and then follow this standard. Some ongoing efforts, such as ORCID, Scopus Author Identifier or Web of Science Researcher Id might eliminate this issue (Arunachalam & Madhan, 2016; LaFlamme, 2016).

Interdisciplinary work: Many existing data sets that are used for AND techniques belong to Computer Science. But, in this contemporary era, the boundaries between different knowledge areas are blurring and interdisciplinary research is gaining popularity. Every discipline has its own pattern of a number of co-authors like Medical Sciences have a lot of publications with sole authors, whereas Physics have publications with a lot of co-authors. As recently reported in Nature news that 'Physics paper sets record with more than 5,000 authors'[13]. Multidisciplinary articles authored by many persons from multiple institutions (nationwide or worldwide) may cause 'multiple entity disambiguations' problems.

Incremental disambiguation: It is constantly needed to insert new citation records in disambiguated DL and each reference should be assigned to their respective 'true' authors. One possible solution to this problem could be to re-run disambiguation technique after every insertion of new citation records into the disambiguated DL. This solution is not possible due to two reasons: first, DL usually composed of thousands, sometimes even in several millions of records. Second, this could be applied only in the case of unsupervised techniques. It is not possible to apply this strategy in supervised techniques because every time, we cannot retrain the classifier.

New authors: A huge number of new researchers present their work in different venues. Nearly all AND techniques use static data or existing data. No one has considered the real world situation in which many new authors are publishing. AND techniques should be capable of doing correct predictions about these new authors publications that have not yet published any work (Ferreira *et al.*, 2015). So, there is also a room for working on this kind of solution that takes into account the new authors situation in DLs. This might be a good future research direction for new AND techniques. A related work on creating a synthetic data set for this situation has been the subject of a study conducted by Ferreira *et al.* (2012) but still this is an approximation to the situation and only creates synthetic data of existing authors in the DL data set.

7 Conclusions

A substantial review of a broad range of existing author name disambiguation techniques has been presented, compared and classified into categories on the basis of used techniques. Some representative methods of all categories are included in this survey. It was observed in the surveyed techniques that the vast majority of the techniques resolved the ambiguity of author names by using some predefined similarity measures or by proposing new similarity functions and then clustering the results of these similarities in citations data. Some methods disambiguate by using graph-based or community detection-based techniques. These AND techniques have been classified in to five categories: supervised, unsupervised, semi-supervised, graph-based and heuristic-based. Supervised techniques have usually better performance on their training data set than other technique but they require costly training data for each class. These techniques required that complete representation data of all target classes should be present in the training data set. Selection of suitable similarity measures and clustering technique in unsupervised techniques is a difficult task. Semi-supervised works well in the case of a limited number of target ambiguous authors but fails when the number of ambiguous authors increases. Graph-based methods are new and need some maturity and rigorous proof. Heuristic-based solutions not always work; sometimes they give very bad results than other techniques.

Some open challenges and future research directions that need some more attention from researchers have been discussed. The new and sole authors cases in citations were reviewed in the context of AND.

[13] http://www.nature.com/news/physics-paper-sets-record-with-more-than-5-000-authors-1.17567

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Similarly, the need for a standardized name for publishers is identified. Scalability and interdisciplinary research issues have been examined in detail.

During this review, we identified that we cannot directly compare the performance of one method to other due to several reasons. One most important reason is there is no standard data set available on which all techniques can be compared, some used DBLP, others used Arnetminer, and still, others used MAS or CiteSeer or BDBCOMP data set. Finally, some used their own defined, like Chinese, Korean or German data sets. Similarly, everyone used different performance metrics for comparisons. The majority of studies used a data set of Computer Science field, whereas now research trends have been revolutionized and new mixed fields are emerging that made author name ambiguity problem even harder than it was in the previous decade. This survey will definitely provide new avenues for researchers to advance and discover some AND techniques.

Acknowledgment

The first author is partially supported by a grant of the Higher Education Commission (HEC), Pakistan.

References

Amancio, D. R., Oliveira, O. N. Jr & Costa, L. D. F. 2015. Topological-collaborative approach for disambiguating authors names in collaborative networks. *Scientometrics* **102**(1), 465–485.

Arunachalam, S. & Madhan, M. 2016. Adopting orcid as a unique identifier will benefit all involved in scholarly communication. *The National Medical Journal of India* **29**(4), 227–234.

Aswani, N., Bontcheva, K. & Cunningham, H. 2006. Mining information for instance unification. In *International Semantic Web Conference*, 329–342. Springer.

Bekkerman, R. & McCallum, A. 2005. Disambiguating web appearances of people in a social network. In *Proceedings of the 14th International Conference on World Wide Web*, 463–470. ACM.

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

Carrasco, R. C., Serrano, A. & Castillo-Buergo, R. 2016. A parser for authority control of author names in bibliographic records. *Information Processing & Management* **52**(5), 753–764.

Chin, W.-S., Zhuang, Y., Juan, Y.-C., Wu, F., Tung, H.-Y., Yu, T., Wang, J.-P., Chang, C.-X, Yang, C.-P., Chang, W.-C. Huang, K.-H., Kuo, T.-M., Lin, S.-W., Lin, Y.-S., Lu, Y.-C., Su, Y.-C., Wei, C.-K., Yin, T.-C., Li, C.-L., Lin, T.-W., Tsai, C.-H., Lin, S.-D., Lin, H.-T. & Lin, C.-J. 2014. Effective string processing and matching for author disambiguation. *The Journal of Machine Learning Research* **15**(1), 3037–3064.

Chisholm, A. & Hachey, B 2015. Entity disambiguation with web links. *Transactions of the Association for Computational Linguistics* **3**, 145–156.

Christen, P. 2006. A comparison of personal name matching: techniques and practical issues. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, 290–294. IEEE.

De Carvalho, A. P., Ferreira, A. A., Laender, A. H. & Gonçalves, M. A. 2011. Incremental unsupervised name disambiguation in cleaned digital libraries. *Journal of Information and Data Management* **2**(3), 289.

Elliott, S. 2010. Survey of author name disambiguation: 2004 to 2010. *Library Philosophy and Practice* **473**, `http://digitalcommons.unl.edu/libphilprac/473/`.

Esperidião, L. V. B., Ferreira, A. A., Laender, A. H., Gonçalves, M. A., Gomes, D. M., Tavares, A. I. & de Assis, G. T. 2014. Reducing fragmentation in incremental author name disambiguation. *Journal of Information and Data Management* **5**(3), 293.

Fan, X., Wang, J., Pu, X., Zhou, L. & Lv, B. 2011. On graph-based name disambiguation. *Journal of Data and Information Quality (JDIQ)* **2**(2), 10.

Ferreira, A. A., Gonçalves, M. A. & Laender, A. H. 2012. A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record* **41**(2), 15–26.

Ferreira, A. A., Gonçalves, M. A. & Laender, A. H. 2015. Automatic methods for disambiguating author names in bibliographic data repositories. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 297–298. ACM.

Ferreira, A. A., Veloso, A., Gonçalves, M. A. & Laender, A. H. 2010. Effective self-training author name disambiguation in scholarly digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, 39–48. ACM.

Ferreira, A. A., Veloso, A., Gonçalves, M. A. & Laender, A. H. 2014. Self-training author name disambiguation for information scarce scenarios. *Journal of the Association for Information Science and Technology* **65**(6), 1257–1278.

Giunchiglia, F. & Shvaiko, P. 2003. Semantic matching. *The Knowledge Engineering Review* **18**(3), 265–280.

Gurney, T., Horlings, E. & Van Den Besselaar, P. 2012. Author disambiguation using multi-aspect similarity indicators. *Scientometrics* **91**(2), 435–449.

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Han, D., Liu, S., Hu, Y., Wang, B. & Sun, Y. 2015. Elm-based name disambiguation in bibliography. *World Wide Web* **18**(2), 253–263.

Han, H., Giles, L., Zha, H., Li, C. & Tsioutsiouliklis, K. 2004. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 2004 joint ACM/IEEE conference on Digital Libraries, 2004*, 296–305. IEEE.

Han, H., Xu, W., Zha, H. & Giles, C. L. 2005. A hierarchical naive bayes mixture model for name disambiguation in author citations. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, 1065–1069. ACM.

Huynh, T., Hoang, K., Do, T. & Huynh, D. 2013. Vietnamese author name disambiguation for integrating publications from heterogeneous sources. In *Asian Conference on Intelligent Information and Database Systems*, 226–235. Springer.

Imran, M., Gillani, S. & Marchese, M. 2013. A real-time heuristic-based unsupervised method for name disambiguation in digital libraries. *D-Lib Magazine* **19**(9), 1.

Johnson, D. B. 1975. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing* **4**(1), 77–84.

Kofod-Petersen, A. 2012. How to do a structured literature review in computer science. Document released as a guide to performing a Structured Literature Review at NTNU. `https://pdfs.semanticscholar.org/f9e7/b1f645ddeddfbf702558f554dd316a7692ae.pdf`.

Krzywicki, A., Wobcke, W., Bain, M., Martinez, J. C. & Compton, P. 2016. Data mining for building knowledge bases: techniques, architectures and applications. *Knowledge Engineering Review* **31**(2), 97–123.

Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M. K. & Ahalt, S. 2014. Privacy preserving interactive record linkage (ppirl). *Journal of the American Medical Informatics Association* **21**(2), 212–220.

LaFlamme, M. 2016. On the problem of the namesake. *Cultural Anthropology* **31**(1), 1–3.

Lee, D., Kang, J., Mitra, P., Giles, C. L. & On, B.-W. 2007. Are your citations clean? *Communications of the ACM* **50**(12), 33–38.

Levin, F. H. & Heuser, C. A. 2010. Evaluating the use of social networks in author name disambiguation in digital libraries. *Journal of Information and Data Management* **1**(2), 183.

Levin, M., Krawczyk, S., Bethard, S. & Jurafsky, D. 2012. Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology* **63**(5), 1030–1047.

Liu, Y., Li, W., Huang, Z. & Fang, Q. 2015. A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology* **66**(3), 634–644.

Liu, Y. & Tang, Y. 2015. Network based framework for author name disambiguation applications. *International Journal of u-and e-Service, Science and Technology* **8**(9), 75–82.

Maguire, E. J. 2016. Ethnicity sensitive author disambiguation using semi-supervised learning. In *Proceedings of the Knowledge Engineering and Semantic Web: 7th International Conference*, KESW 2016 **649**, 272. Springer, 21–23 September 2016.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of Internal Medicine* **151**(4), 264–269.

Murnane, E. L., Haslhofer, B. & Lagoze, C. 2013. Reslve: leveraging user interest to improve entity disambiguation on short text. In *Proceedings of the 22nd International Conference on World Wide Web*, 1275–1284. ACM.

Nicholson, S. W. & Bennett, T. B. 2016. Dissemination and discovery of diverse data: do libraries promote their unique research data collections? *International Information & Library Review* **48**(2), 85–93.

On, B.-W., Elmacioglu, E., Lee, D., Kang, J. & Pei, J. 2006. Improving grouped-entity resolution using quasi-cliques. In *Sixth International Conference on Data Mining (ICDM'06)*, 1008–1015. IEEE.

On, B.-W., Lee, D., Kang, J. & Mitra, P. 2005. Comparative study of name disambiguation problem using a scalable blocking-based framework. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, 344–353. ACM.

On, B.-W., Lee, I. & Lee, D. 2012. Scalable clustering methods for the name disambiguation problem. *Knowledge and Information Systems* **31**(1), 129–151.

Onodera, N., Iwasawa, M., Midorikawa, N., Yoshikane, F., Amano, K., Ootani, Y., Kodama, T., Kiyama, Y., Tsunoda, H. & Yamazaki, S. 2011. A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search. *Journal of the American Society for Information Science and Technology* **62**(4), 677–690.

Oramas, S., Espinosa-Anke, L., Sordo, M., Saggion, H. & Serra, X. 2016. Elmd: an automatically generated entity linking gold standard dataset in the music domain. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC*.

Palfrey, J. 2016. Design choices for libraries in the digital-plus era. *Daedalus* **145**(1), 79–86.

Peng, H.-T., Lu, C.-Y., Hsu, W. & Ho, J.-M. 2012. Disambiguating authors in citations on the web and authorship correlations. *Expert Systems with Applications* **39**(12), 10521–10532.

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H., Gonçalves, M. A. & Ferreira, A. A. 2009. Using web information for author name disambiguation. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, 49–58. ACM.

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press

Provost, F. & Kohavi, R. 1998. Guest editors' introduction: on applied research in machine learning. *Machine Learning* **30**(2), 127–132.

Pyle, R. L. 2016. Towards a global names architecture: the future of indexing scientific names. *ZooKeys* **550**, 261–281.

Santana, A. F., Gonçalves, M. A., Laender, A. H. & Ferreira, A. A. 2015. On the combination of domain-specific heuristics for author name disambiguation: the nearest cluster method. *International Journal on Digital Libraries* **16**(3–4), 229–246.

Scholtes, J. C. & Maes, F. P. E., *et al*. 2016. System and method for authorship disambiguation and alias resolution in electronic data. US Patent 9,264,387.

Schulz, C., Mazloumian, A., Petersen, A. M., Penner, O. & Helbing, D. 2014. Exploiting citation networks for large-scale author name disambiguation. *EPJ Data Science* **3**(1), 1.

Seol, J.-W., Lee, S.-H. & Kim, K.-Y. 2016. Author disambiguation using co-author network and supervised learning approach in scholarly data. *International Journal of Software Engineering and Its Applications* **10**(4), 73–82.

Shin, D., Kim, T., Choi, J. & Kim, J. 2014. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics* **100**(1), 15–50.

Song, Y., Huang, J., Councill, I. G., Li, J. & Giles, C. L. 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 342–351. ACM.

Tang, J., Fong, A. C., Wang, B. & Zhang, J. 2012. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering* **24**(6), 975–987.

Tang, L. & Walsh, J. P. 2010. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics* **84**(3), 763–784.

Torvik, V. I. & Smalheiser, N. R. 2009. Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **3**(3), 11.

Tran, H. N., Huynh, T. & Do, T. 2014. Author name disambiguation by using deep neural network. In *Asian Conference on Intelligent Information and Database Systems*, 123–132. Springer

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F. & Pinheiro, D. 2012. A boosted-trees method for name disambiguation. *Scientometrics* **93**(2), 391–411.

Wang, P., Zhao, J., Huang, K. & Xu, B. 2014. A unified semi-supervised framework for author disambiguation in academic social network. In *International Conference on Database and Expert Systems Applications*, 1–16. Springer.

Wang, X., Tang, J., Cheng, H. & Philip, S. Y. 2011. Adana: active name disambiguation. In *2011 IEEE 11th International Conference on Data Mining*, 794–803. IEEE.

Weiss, A. 2016. Examining massive digital libraries (mdls) and their impact on reference services. *The Reference Librarian* **57**(4), 286–306.

Wu, H., Li, B., Pei, Y. & He, J. 2014. Unsupervised author disambiguation using Dempster-Shafer theory. *Scientometrics* **101**(3), 1955–1972.

Zhao, J., Wang, P. & Huang, K. 2013. A semi-supervised approach for author disambiguation in KDD CUP 2013. In *Proceedings of the 2013 KDD CUP 2013 Workshop*, 10. ACM.

Zhu, J., Yang, Y., Xie, Q., Wang, L. & Hassan, S.-U. 2014. Robust hybrid name disambiguation framework for large databases. *Scientometrics* **98**(3), 2255–2274.

Zhu, L., Ghasemi-Gol, M., Szekely, P., Galstyan, A. & Knoblock, C. A. 2016. Unsupervised entity resolution on multi-type graphs. In *International Semantic Web Conference*, 649–667. Springer.

Zhu, Y. & Li, Q. 2013. Enhancing object distinction utilizing probabilistic topic model. In *2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia)*, 177–182. IEEE.

https://doi.org/10.1017/S0269888917000182 Published online by Cambridge University Press