Role of References in Similarity Estimation of Publications

Muhammad Shoaib[1], Ali Daud[2], and Malik Khiyal[3]
[1,2]Department of Computer Science and Software Engineering, International Islamic University, Pakistan
[1]Department of Computer Science, COMSATS Institute of Information Technology, Pakistan
[3]Faculty of Computer Sciences, Preston University Islamabad, Pakistan

Abstract: *Similarity estimation among publications is very important in classification and clustering techniques for grouping, indexing, citation matching and Author Name Disambiguation (AND) purposes. Publication attributes are basic sources of information and play important role in similarity estimation. Most of the works in AND use title, co-authors and venue attributes for estimating similarity among publications. Many other sources of information such as self-citations, shared citations and references, topic of the publications and abstracts have also been employed to estimate optimal similarity among publications. Recently, in the field of Academic Document Clustering (ADC), reference marker contexts have been utilized for this purpose. However, the use of citations and references is less common since only a few databases include this information. In this paper, we propose to use two components of references (co-authors and titles of references) as sources of information and investigate the importance of these components in similarity estimation. To the best of our knowledge, this is the first endeavour to exploit components of references as sources of information. Experiments conducted on real publication datasets reveal that these components of references are significant source of information for similarity estimation among publications.*

Keywords: *AND, references, vector space model, cosine similarity, citation matching.*

Received May 16, 2014; accepted September 11, 2014; published online July 11, 2015

1. Introduction

Similarity estimation among research publications is very important in text mining tasks. Publication attributes are basic source of information and play important role in similarity estimation. Digital Library (DL) related tasks like Academic Document Clustering (ADC)/ classification, citation mining and Author Name Disambiguation (AND) exploit different types of information (publication attributes) to estimate optimal similarity among publications. A publication has attributes such as title, co-authors, venue (name of journal, conference, etc.,), year of publication, abstract, key words, citations and references. Researchers in DL community have exploited them in various tasks. Out of these attributes title, co-authors and venue are considered the most important ones; and we, in this work, refer them as triplet attributes or simply triplets. Beside the above mentioned publication attributes AND works exploit several other sources of information such as user feedback [5, 31], topics of publication [2] information from the web [31]. Although, researchers have exploited variety of local (title, venue, etc.,) and global (topics) sources of information yet they have focused the references a little.

Exploiting more and more attributes increases the execution cost of text mining tasks. To minimize the cost researchers exploit necessary and informative attributes. For example, in ADN works and citation matching techniques exploiting triplets is very common practice [6]. Almost half of AND works use only these three attributes [6]. Using title and venue attributes to estimate publications similarity may not be real picture of their similarities. Two publications having totally different titles or venues may belong to the same topic(s) and on the other hand, two publications having high title or venue similarity may belong to two different areas as the title and venue attributes face scarcity of words problem. Words scarcity problem means that a title or venue has only few words to represent the topic(s) of a publication or research area of a venue.

To overcome words scarcity problem, we propose to employ components of references as sources of information for similarity estimation. It can be argued that words scarcity problem may be resolved by comparing complete scripts. This solution is too much time consuming and not scalable. Further, complete scripts are not freely and easily available in many databases. On the other hand, references are easily available from almost all Bibliographic Databases (BDs).

1.1. Terminology

- *Publication*: It refers to any published literary work like research paper, book chapter and report. Publication, research publication, paper and academic document have been used interchangeably.
- *Citation*: The complete reference to a publication in a bibliographic database. It usually contains names

of co-authors, title, venue and year of publication, etc.

- *References*: The bibliographic list given at the end of a publication.
- *Ref-Titles*: All titles of references of a publication are combined and we name it references titles or ref-titles.
- *Ref-Co-Authors*: All co-authors of references of a publication are combined together and we name it references co-authors or ref-co-authors for short.
- *Document*: The word document has no specific meanings in this work. It means any text document. We, at some occasions, use this term to generalize the discussion. So, document may mean a citation or a publication or even a text string.

Co-authors, title and venue are citation attributes as well as publication attributes. Ref-titles and ref-coauthors are not part of citations hence they are not citation attributes. They are part of publications and thus are referred as publication attributes. All citation attributes are also publication attributes but vice versa is not true.

In this work, we investigate the importance of two components of references in similarity estimation of publications. It is our hypothesis that similarity among references of two publications is relatively closer to the actual value than the similarities among their titles or venues. The term actual value means the similarity calculated by comparing complete scripts of publications.

- *Contribution*: To the best of our knowledge, it is the first work which focuses to utilize components of references (ref-titles and ref-coauthors) for similarity estimation of publications. We estimate pair-wise similarity of publication attributes to analyse the impact of ref-titles and ref-coauthors in similarity estimation.

Experiments on real publication datasets reveal that ref-titles and ref-coauthors are reliable sources of information for similarity estimation of publications.

Rest of the paper is organized as: Section 2 covers related work. Section 3 describes problem statement. Section 4 presents the proposed solution. Section 5 discusses the results. Section 6 summarises our work and with the description of future directions.

2. Related Works

Publication attributes are basic source of information and play important role in similarity estimation. Most of the works in AND like Han *et al.* [10] use triplet attributes for estimating similarity among publications. Almost half of AND works use only triplet attributes [6]. Works in AND exploit diverse types of attributes such as self citation [14, 27, 32], abstract [27, 32], user feedback [5, 31] topic of the publication [2, 10, 24, 26, 31], author affiliation [32], authors email addresses [32] and web information [31]. Shu *et al.* [24] use latent dirichlet allocation [3] for topic modelling [4]. Kleb and Volz [13] use ontological or semantic techniques [23] for guessing topics of publications. Torvik *et al.* [29] use eight different attributes. Smalheiser and Torvik [25] enhance their task of [29] by including first name and its variants, emails and correlations between last names and affiliation words.

In the field of ADC, use of reference markers and their contexts have gained much attention [15, 16]. In works of Mercer and Marco [15] text surrounding a reference marker is extracted to determine the relatedness between two publications connected by that reference marker. Aljaber *et al.* [1] use contexts of reference markers to optimize similarity among publications. Jeon [12] crawls the comments related to the papers cited in the related works sections and then provides useful information regarding the cited papers and how much similar are the cited papers and the paper that is citing those papers.

Levin *et al.* [14, 27, 32] use self-citations to investigate whether the citing and cited publications belong to the same author. They consider two papers authored by the same person if one of them cites the other. We on the other hand, compare ref-titles and ref-coauthors of all references. So, our work is totally different from their work. Aljaber *et al.* [1] exploit reference markers contexts to estimate similarity between two publications. Their approach scans the whole script to find reference markers contexts. These contexts are then compared to estimate the similarity. Their work is different from ours that they compare reference markers contexts while we compare the ref-titles and ref-coauthors of references. The reference markers contexts may or may not represent the cited work properly as every writer describes the cited work in his/her own style and according to the flow and need of the paper. Two reference marker contexts of the same work by two different authors may have totally different wordings. The closest works (for estimation of publications similarity) to our’s are Schulz *et al.* [16, 23]. Schulz *et al.* [18, 28] exploit shared citations (citing papers) and shared references whereas Tang and Walsh [28] use only shared references. Contrary to them we exploit all the co-authors and titles of all references because two non common references of two publications may have few co-authors and/or title words in common. To the best of our knowledge this is the first work that uses two components of complete list of references for estimating publications similarity. Ref-venues can also be investigated whether this attribute is a good source of information or not.

3. Problem Statement

Many techniques of AND and ADC use only triplets to measure similarity between publications. Triplets specially title and venue attributes face words scarcity problem. Ref-titles may resolve words scarcity problem as this attribute has many words related to the publication. Co-authors attribute is considered very powerful source of information for grouping the publications of the same author. It is assumed that the co-authors attribute is the least variant in publications of the same author. It is further assumed that co-authors of an author are usually changed when he/she changes research topic. This assumption may not be true for each author. Usually, a researcher at university works with different students. Every year new students join and the previous leave his/her group. In these situations it becomes difficult to group the publications of an author or publications of the same topic on the base of co-authors attribute. Ref-coauthors may be useful source of information in such scenarios because ref-coauthors attribute usually consists of many collaborative research groups (co-authors) working on the topic(s) of the publication.

We, in this paper, consider references similarity in two ways: Whether ref-co-authors and ref-titles similarities are closer to or farther than actual similarity than those of title and co-authors'; and whether ref-co-authors and ref-titles help improve text mining tasks or not.

4. Proposed Solution

For title, ref-titles, venue and complete script, state of the art cosine similarity representing the publications in VSM [17] is used. For co-authors and ref-coauthors we use our own measures proposed in an unpublished work [22].

4.1. Similarity Measure for Title, Ref-titles and Complete Script

We combine all titles of all references of a publication into one title and name it as ref-titles. If there are r references of a publication p then there are r titles as each reference has exactly one title. Aggregating r titles into one title gives us one ref-title. The term "Ref-titles" is considered singular. The ending "s" of "ref-titles" represents that there are r titles present in r references of one publication.

We use cosine similarity as it is the most popular measure [21, 30] for estimating document similarity based on VSM. The similarity between two documents *a* and *b* can be defined as the normalized inner product of the two corresponding vectors *a* and *b*[1].

$$Sim_{cos}(a,b) = \frac{a.b}{|a| \times |b|} = \frac{\sum_{t\in(a\cap b)} (w_{a,t} \times w_{b,t})}{\sqrt{\sum_{t\in a} w_{a,t}^{2} \times \sum_{t\in b} w_{b,t}^{2}}} \quad (1)$$

Where $(a \cap b)$ represents common terms of documents *a* and *b*; $w_{a,\ t}$ and $w_{b,\ t}$ are the weights of term *t* in documents *a* and *b* respectively.

4.2. Similarity Measure for Co-Authors and Ref-Co-authors

Like ref-titles, we combine all co-authors lists of all references of a publication into one co-authors list and name it as ref-coauthors. If there are r references of a publication p then there are r co-authors lists as each reference has exactly such list. Aggregating r co-authors lists into one list gives us one ref-coauthors list. Like ref-titles, the term "Ref-co-authors" is also, considered singular.

Cosine function can be applied to co-authors attribute where variations in names are minimal. It is not a better solution for entity names where a name has variant forms especially when a name has multiple tokens. For example, "Muhammad Shoaib Kamboh" can be written in many ways like: "M. S. Kamboh", M. Shoaib Kamboh, etc., cosine function considers each variant form of a token as different term. To estimate similarity between two names n_i and n_j we exploit jaccard like formula proposed in our unpublished work[2]. This is given in Equation 2:

$$Sim_{nam}(n_i, n_j) = \frac{e*\alpha + b*\beta + q*\gamma}{z*0.5 + h*100} * \log(z+2) \quad (2)$$

Where *α*, *β* and *γ* represent weights of *e*, *b* and *q* respectively; *e* represents number of exact matching tokens[3], *b* abbreviation matching tokens, *q* abbr-initial matching tokens; *h*, number of conflicting tokens; and *z*, total number of tokens in both names. In above equation *h**100 factor decreases similarity value of two different names (having conflicting tokens) near to 0. Why we assign different weights to different types of tokens is discussed in Appendix A.

To estimate co-authors and ref-coauthors similarity we exploit simple jaccard formula given in Equation 3:

$$Sim_{CA}(a,b) = \frac{2*(\Gamma)}{N} \quad (3)$$

Where *N* is the total number of names in both publications and *Γ* is the number of names having Sim_{nam}>threshold. Sim_{nam} is estimated through Equation 2. Above equation gives co-authors and ref-coauthors similarity between two publications *a* and *b*.

5. Results and Discussion

In this section we explain the results generated on real publication datasets. We performed experiments on two types of datasets: I.e., publication datasets of ambiguous authors and publication dataset of different subjects. We collected six publication datasets of different ambiguous authors as exploited by different works like [7, 11]. We included only those ambiguous names and individual authors for whose publications we could collect the references along with other citation attributes. In our experimental datasets, each

[1] Bold face letters represent vector form of a document.

[2] Different types of tokens mentioned here are defined in Appendix A

ambiguous dataset contains 44-150 records and 3-6 individual authors. Table 1 shows statistics of six datasets. We performed stemming and stop words removal as preprocessing steps for title and ref-titles attributes.

Table 1. Publication datasets of ambiguous authors.

Ambiguous Names	No. of Records	No. of Authors	Ambiguous Names	No. of Records	No. of Authors
Ajay Gupta	134	6	Hui Fang	87	4
Bing Liu	105	5	Jim Smith	44	3
Cheng Chang	61	4	Rakesh Kumar	150	6

We divide each ambiguous dataset into sub-datasets in such a way that each sub-dataset contains records of one and only one individual author. This technique results into twenty eight sub-datasets.

The pair-wise attribute similarity between each pair of records of each sub-dataset has been computed. Main focus is to analyse whether references attributes help improve publications similarity or not. The results are given in Tables 2 and 3.

Table 2. Comparison between similarity values of title and ref-titles attributes.

Ambiguous Name	Intra Sub-datasets Title Sim Avg.	Time Consumed (sec.)	Intra Sub-Datasets Ref-Titles Sim Avg.	Time Consumed (sec.)
Ajay Gupta	0.033946824	0.7644013	0.061333268	1.3572023
Bing Liu	0.024543994	0.670203	0.058223043	1.2932041
Cheng Chang	0.060755893	0.6096011	0.079818487	0.9204017
Hui Fang	0.044190486	0.6396011	0.076130073	1.1204016
Jim Smith	0.055144755	0.4212007	0.07249837	0.8112015
Rakesh Kumar	0.025779532	0.9360016	0.063823024	1.8096032
Total	**0.244361485**	**4.0410088**	**0.411826264**	**7.3120144**

Table 3. Comparison between similarity values of co-authors and ref-coauthors attributes.

Ambiguous Name	Intra Sub-Datasets Avg. Co-auths Sim.	Time Consumed (sec.)	Intra Sub-Datasets Avg. Ref-Coauthors Sim.	Time Consumed (sec.)
Ajay Gupta	0.154066219	1.1870165	0.240939995	6.2480121
Bing Liu	0.117684718	1.2840735	0.180463357	9.5620231
Cheng Chang	0.464357143	1.4570832	0.302015341	1.6700025
Hui Fang	0.317898957	1.5990917	0.175691372	3.2340073
Jim Smith	0.420530456	1.3640782	0.174297317	4.4920047
Rakesh Kumar	0.329393691	1.5520035	0.173592064	8.5570157
Total	**1.803931184**	**8.4433466**	**1.246999445**	**32.7630654**

Table 2 shows comparison between similarity values of title and ref-titles attributes. The second column i.e., "Intra Sub-datasets Title Sim Avg." reports average title similarity between the records of a sub-dataset excluding self-comparisons. Forth column reports the same thing for ref-titles attribute. Third and fifth columns show the time consumed in seconds to estimate respective attribute similarity values for intra sub-dataset records.

Table 2 shows that ref-titles similarity is always higher than title similarity. On the average ref-titles similarity is almost 1.7 times higher than title similarity. Estimating ref-titles similarity is comparatively more time consuming than estimating title similarity. On the average time consumed to calculate ref-titles similarity is almost 1.8 times greater than the time consumed for estimating title similarity. The disadvantage of greater time consumption is negligible as compared to the advantage of similarity information from ref-titles attribute. Table 2 shows that ref-titles similarity is more reliable source of information for publications datasets of ambiguous authors. Greater values of ref-titles attribute guarantees that it can be used as additional source of information in AND process.

Table 3 is similar to Table 2 with only difference that it shows similarity values for co-authors and ref-co-authors attributes.

Table 3 shows that for some datasets (e.g., Ajay Gupta) ref-coauthors similarity is higher than co-authors similarity and for some datasets (e.g., Jim Smith) situation is reverse. For example, ref-coauthors similarity is 1.56 times of co-authors similarity for Ajay Gupta dataset and 0.41 times for Jim Smith's dataset. On the average co-authors similarity is almost 1.45 times higher than ref-coauthors similarity. Estimating ref-coauthors similarity is more time consuming than estimating coauthors similarity. On the average time consumed to calculate ref-coauthors similarity is almost 4.0 (3.88) times greater than the time consumed for coauthors similarity. The disadvantage of additional time consumption is bearable. In trade of CPU time cost we get an additional source of information. Table 3 reveals that although ref-co-authors attribute is not as powerful source of information as co-authors attribute yet it is useful source of information for publications datasets of ambiguous authors.

Now, let us analyse whether ref-titles and ref-coauthors similarity is closer to actual similarity than title, co-authors and venue similarity or not. We have prepared three small datasets of almost 30 publications from three different subjects. Each tiny dataset contains publications of the same topic from respective subject. These datasets are not from the same author or same ambiguous name instead they are from the same topic. For these datasets, title, ref-titles, co-authors, ref-coauthors, venues and complete script similarities have been estimated. The results are shown in Table 4. Similarity values for title, ref-titles, venues and complete scripts mentioned in Table 4 are estimated through cosine representing documents in VSM; for names, Equation 2 is used; and for co-authors and ref-co-authors Equation 3 is employed.

Table 4 shows ref-titles similarity is the closet to actual similarity and it is almost 3 times higher than title similarity. It is clear that ref-titles are good source of information for topic based publications datasets. While analysing ref-titles and ref-coauthors attributes of publications we get some interesting pieces of information.

Table 4. Comparison between similarity values of title, ref-titles, co-authors, ref-coauthors and venue attributes w.r.t. actual similarity (complete script sim).

Datasets	Title Sim	Ref-Titles Sim	Co-Authors Sim	Ref-Coauthors Sim	Venue Sim	Complete Script Sim
Computer Sc.	0.046	0.111	0.015	0.027	0.017	0.304
Physics	0.028	0.160	0.013	0.012	0.002	0.155
Economics	0.031	0.052	0.004	0.002	0.001	0.121
Total Sim	0.105	0.323	0.032	0.041	0.020	0.580
Average Sim	**0.035**	**0.108**	**0.011**	**0.014**	**0.007**	**0.19**

While analysing ref-titles and ref-coauthors attributes of publications we get some interesting pieces of information.

- Consider two publications of Shan *et al*. [19, 20] having same title and same coauthors published in two different venues. Their title and coauthor similarity is 1.00 and venue similarity is 0.034. Title and coauthors similarity reveal that these two publications are not different publications while venue similarity depicts that they are two different publications. Ref-titles and ref-coauthors similarity values (0.280877 and 0.779 respectively) show that they share a reasonable amount of data. To estimate actual similarity between the two publications we compare their abstracts and then complete scripts. Abstract and complete script similarity values are 0.348094 and 0.544173 respectively. Out of all these values complete script similarity value i.e., 0.544173 is the most reliable and genuine. The average of ref-titles and ref-coauthors similarity i.e., 0.53 is the closet to the actual similarity value.
- Consider two publications of Gupta and Beckmann [8, 9] having same title and coauthors but different venues. Their title and coauthor similarity is 1.00 and venue similarity is 0.073. Title and coauthors similarity reveal that these two publications are not different publications while venue similarity depicts that they may share little amount of data. Ref-titles and ref-coauthors similarity values (0.912 and 1.00 respectively) show that they share almost whole text. To estimate actual similarity picture between the two publications we compare their complete scripts. Actual similarity value is also equal to 1.0. Out of all these values full script similarity value i.e., 1.0 is the most reliable and genuine. In this case the similarity values of all attributes except venue are very close or equal to the actual value. After getting high similarity value of ref-titles and ref-coauthors we manually investigated the two publications. Our notion was that they would be the exact copy of each other. After investigation it proved that our notion was absolutely true.

From above discussion, it is concluded that if two publications have same titles and co-authors but vary in their references then one of them may be the extension of the other. Also, if two publications have same titles, co-authors and references then it is quite possible that they are copy of each other. From this discussion it reveals that references attributes help a large in certain situations to decide whether two documents are copy of each other or not. This simple test may also be performed to help decide the plagiarism process.

We were interested to figure out whether two main components of references (ref-titles and ref-coauthors) could be used as sources of information to improve text mining tasks that rely on publication similarity. We started this research with the notion that authors include those references which relate to the topic(s) of the publication. Our notion is logical and has been proved by empirical results generated from real life datasets. Above results and discussion show that ref-titles and ref-coauthors improve publication similarity. Text mining tasks like ADC, citation matching and AND base on publication similarity. So, we can claim that ref-titles and ref-coauthors can be helpful in such text mining tasks. References components will surely improve accuracy of ADC, AND process and many other tasks which rely on document similarity. They provide reliable information about the amount of data the two documents share with each other. Ref-titles attribute is more reliable as compared to the title attribute. Ref-coauthors attribute though not more informative than co-authors attribute yet provides reasonable amount of similarity information. From this discussion we conclude that our proposed idea of exploiting references for estimating academic documents similarity is worthwhile.

6. Conclusions and Future Works

In this paper, it is proved that ref-titles and ref-coauthors attributes help improve publications similarity. Extensive experiments have been performed on publication datasets of ambiguous authors and publication datasets having same topic. From experiments, it is concluded that references attributes provide good source of similarity information for publications. Ref-titles attribute is more reliable than title attribute. Ref-coauthors attribute though not more informative than co-authors attribute yet provides reasonable amount of similarity information. From this discussion, it is concluded that the proposed idea of exploiting two main components of references for estimating academic documents similarity is worthwhile. As future directions we will employ the same methodology to analyse the amount of precision and recall improved in AND process and in ADC collecting larger datasets. We are also interested to employ ref-venue attribute to analyse its impact on publication similarity.

References

[1] Aljaber B., Stokes N., Bailey J., and Pei J., "Document Clustering of Scientific Texts using Citation Contexts," *Information Retrieval*, vol. 13, no. 2, pp. 101-131, 2010.

[2] Bhattacharya I. and Getoor L., "A Latent Dirichlet Model for Unsupervised Entity Resolution," available at: http://linqs.cs.umd.edu/basilic/web/Publications/2006/bhattacharya:sdm06/bhattacharyasdm06.pdf, last visited 2006.

[3] Blei D., Ng A., and Jordan M., "Latent Dirichlet Allocation," available at: https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf, last visited 2003.

[4] Daud A., Li L., and Muhammad F., "Knowledge Discovery through Directed Probabilistic Topic Models, a Survey," *Frontiers of Computer*

Science in China, vol. 4, no. 2, pp. 280-301, 2010.

[5] Fan X., Wang J., Pu X., Zhou L., and Lv B., "On Graph-based Name Disambiguation," *ACM Journal of Data and Information Quality*, vol. 2, no. 2, pp. 1-23, 2011.

[6] Ferreira A., Gonçalves M., and Laender A., "A Brief Survey of Automatic Methods for Author Name Disambiguation," *ACM SIGMOD Record*, vol. 41, no. 2, pp. 15-26, 2012.

[7] Ferreira A., Velosol A., Gonçalves M., and Laender A., "Effective Self-Training Author Name Disambiguation in Scholarly Digital Libraries," *in Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries*, Gold Coast, Australia, pp. 39-48, 2010.

[8] Gupta A. and Beckmann B., "PANSY: A Portable Autonomous Irrigation System," available at: http://www.iasri.res.in/icsi2006/ theme3/ajay.pdf, last visited 2006.

[9] Gupta A. and Beckmann B., "PANSY: A Portable Autonomous Irrigation System," *Journal of Indian Society of Agricultural Statistics*, vol. 61, no. 2, pp. 156-163, 2007.

[10] Han H., Giles L., Zha H., Li C., and Tsioutsiouliklis K., "Two Supervised Learning Approaches for Name Disambiguation in Author Citations," *in Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries*, Tucson, USA, pp. 296-305, 2004.

[11] Han H., Zha H., and Giles L., "Name Disambiguation in Author Citations using a K-Way Spectral Clustering Method," *in Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries*, Denver, USA, pp. 334-343, 2005.

[12] Jeon H., "A Reference Comments Crawler for Assusting Research Paper Writing," *the International Arab Journal of Information Technology*, vol. 11, no. 5, pp. 493-499, 2014.

[13] Kleb J. and Volz R., "Ontology based Entity Disambiguation with Natural Language Patterns," *in Proceedings of the 4th International Conference on Digital Information Management*, Ann Arbor, pp. 1-8, 2009.

[14] Levin M., Krawczyk S., Bethard S., and Jurafsky D., "Citation-based Bootstrapping for Large-Scale Author Disambiguation," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 5, pp. 1030-1047, 2012.

[15] Mercer R. and Marco C., "A Design Methodology for a Biomedical Literature Indexing Tool using the Rhetoric of Science," *in Proceedings of BioLink Workshop in Conjunction with Human Language Technology Conference/ North American Chapter of the Association for Computational Linguistics Annual Meeting*, pp. 77-84, 2004.

[16] Nanba H. and Okumura H., "Towards Multi Paper Summarization using Reference Information," *in Proceedings of the 16th International Joint Conferences on Artificial Intelligence*, pp. 926-931, 1999.

[17] Salton G., Wong A., and Yang C., "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.

[18] Schulz C., Mazloumian A., Petersen A., Penner O., and Helbing D., "Exploiting Citation Networks for Large-Scale Author Name Disambiguation," *EPJ Data Science*, vol. 3, no. 1, pp. 1-14, 2014.

[19] Shan Y., Sawhney H., and Kumar R., "Unsupervised Learning of Discriminative Edge Measures for Vehicle Matching between Non-Overlapping Cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 700-711, 2008.

[20] Shan Y., Sawhney H., and Kumar R., "*Unsupervised Learning of Discriminative Edge Measures for Vehicle Matching between Non-Overlapping Cameras*," *in Proceedings of the Conference on Computer Vision and Pattern Recognition*, Princeton, USA, pp. 894-901, 2005.

[21] Shoaib M., Daud A., and Khiyal M., "An Improved Similarity Measure for Text Documents," *Journal of basic and Applied Scientific Research*, vol. 4, no. 6, pp. 215-223, 2014.

[22] Shoaib M., Daud A., and Khiyal M., "Improving Similarity Measures for Publications with Special Focus on Author Name Disambiguation," *Arabian Journal for Science and Engineering*, vol. 40, no. 6, pp. 1591-1605, 2015.

[23] Shoaib M., Yasin M., Niazi H., Saeed M., and Khiyal S., "Relational WordNet Model for Semantic Search in Holy Quran," *in Proceedings of Internatiojnal Conference on Emerging Technologies*, Islamabad, pakistan, pp. 29-34, 2009.

[24] Shu L., Long B., and Meng W., "A Latent Topic Model for Complete Entity Resolution," *in Proceedings of the 25th IEEE International Conference on Data Engineering*, Shanghai, China, pp. 880-891, 2009.

[25] Smalheiser N. and Torvik V., "Author Name Disambiguation," *Annual Review of Information Science and Technolog*, vol. 43, no. 1, pp. 1-43, 2009.

[26] Song Y., Huang J., and Councill I., "Efficient Topic-based Unsupervised Name Disambiguation," *in Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries*, Vancouver, British Columbia, Canada, pp. 342-351, 2007.

[27] Tang J., Fong A., Wang B., and Zhang J., "A Unified Probabilistic Framework for Name Disambiguation in Digital Library," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 975-987, 2012.

[28] Tang L. and Walsh J., "Bibliometric Fingerprints: Name Disambiguation based on Approximate Structure Equivalence of Cognitive Maps," *Scientometrics*, vol. 84, no. 3, pp. 763-784, 2010.

[29] Torvik V., Weeber M., Swanson D., and Smalheiser N., "A Probabilistic Similarity Metric for Medline Records a Model for Author Name Disambiguation," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 2, pp. 140-158, 2005.

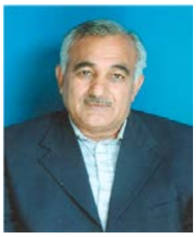
[30] Wan X., "A Novel Document Similarity Measure based on Earth Mover's Distance," *Information Sciences*, vol. 177, no. 18, pp. 3718-3730, 2007.

[31] Wang F., Tang J., Li J., and Wang J., "A Constraint-Based Topic Modeling Approach for Name Disambiguation," *Frontiers of Computer Scienc*e, *China*, vol. 4, no. 1, pp. 100-111, 2010.

[32] Zhang D., Tang J., Li J., and Wang K., "A Constraint-Based Probabilistic Framework for Name Disambiguation," *in Proceedings of the 16th ACM Conference on Information and Knowledge Management*, Lisboa, Portugal, pp. 1019-1022, 2007.

Muhammad Shoaib is currently pursuing PhD degree in computer science from International Islamic University, Islamabad, Pakistan. He received his MSCS degree from the same university and MSc (CS) from University of the Punjab, Lahore, Pakistan. He has more than eight research publications in national and International Journals and National Conferences. His areas of interest are data/text mining, information systems and retrieval and document clustering.

Ali Daud is working as Assistant Professor in the Department of Computer Science at International Islamic University, Islamabad. He obtained his PhD degree from Tsinghua University in 2010. He is head of Data Mining and Information Retrieval Group. He published about 23 papers in reputed International Journals and Conferences. He has taken part in many projects and PI of a project funded by higher education commission, Pakistan. His current research interests include: Text mining, social networks analysis and applications of probabilistic topic models.



Malik Khiyal is currently Professor of Faculty of Computer Science, Preston University, Islamabad. He remained Chairman Department of Computer Sciences and Software Engineering in FJWU Pakistan from 2007 to 2012 and in IIUI, Pakistan from 2002 to 2007. He Served Pakistan Atomic Energy Commission for 25 years (1978-2002) and continuously was involved in different research and development projects of the PAEC. He developed software for underground flow and advanced fluid dynamic techniques. His areas of interest are Numerical Analysis, Analysis of Algorithms, Theory of Automata and Theory of Computation. He has more than 140 research publications. He has supervised four PhD and more than 150 research projects at graduate and postgraduate levels. He is a member of SIAM, ACM, Informing Science Institute, IACSIT. He is associate editor of IJCTE, IJMO, JACN, LNSE and Co-editor of the journals JATIT and International Journal of Reviews in Computing. He is reviewer of Journals, IJCSIT, JIISIT, IJCEE and CEE of Elsevier.

Appendix A

The material given here is taken from reference [22].

Types of Tokens in Names:

- *Full Match Token*: A non-abbreviated token t'_l from name n_i that matches exactly to a non-abbreviated token t'_m from name n_j is considered as full matching token. For example, in Table 5, token t'_1 from name n_1 fully matches to the token t'_2 from name n_4.
- *Abbreviation Match Token*: An abbreviated token t'_l from name n_i that matches exactly to an abbreviated token t'_m from name n_j is referred as abbreviation match token. For example, in Table 5, token t'_1 ("M.") from name n_2 exactly matches to token t'_1 from name n_5.
- *Abbr-Initial Match Token*: An abbreviated token t'_l from name n_i that matches to the initial letter of a non abbreviated token t'_m from name n_j ignoring dot (.) of abbreviated token is considered as abbr-initial match token. For example, in Table 5, token t'_1 of name n_2 ("M.") matches initial letter of t'_1 of n_1 ("Muhammad").
- *Missing Token*: If two names n_i and n_j do not have equal number of tokens then at least one token of n_i or n_j cannot be compared to that of n_j's or n_i's. This is the case of missing token. Consider names n_2 and n_6 in Table 5. Names n_2 and n_6 have two matching tokens but n_2 does not have any token to be compared to the third token (Kamboh) of n_6. Kamboh in n_6 is the missing token.
- *Conflicting Tokens*: If two tokens t'_l and t'_m from two names n_i and n_j do not fall in any of the above categories then the tokens t'_l and t'_m are considered as conflicting tokens. For example, in Table 5, token

t'_2 in mane n_3 does not match to any of the tokens in name n_7. Similarly token t'_2 of mane n_2 does not match to any of the tokens of name n_7. Missing and conflicting tokens are different from each other and they should be treated differently. Missing tokens case occurs only when number of tokens in two names is unequal whereas conflicting tokens case is irrespective of this condition.

Assumption I

The probability that two names (n_i and n_j) sharing full matching tokens belong to the same person is higher than that of sharing abbreviated tokens. Similarly, the probability that two names sharing abbreviated matching tokens belong to the same person is higher than that of sharing abbr-initial tokens.

For example, in Table 5, it is more probable that n_1 and n_4 belong to the same person than the names n_2 and n_5. In n_2 and n_5 “M.” may stand for any token like Mahmood, Mansha, Majid and Maira.

Table 5. Names and notations used for explanation.

Names	Notations	Names	Notations
Muhammad Shoaib	n_1	Shoaib Muhammad	n_4
M. Shoaib	n_2	M. Shoaib	n_5
M. Shoaib kamboh	n_3	M. Shoaib kamboh	n_6
M. Safdar Kamboh	n_7		

Why do we Assign Different Weights to Different Types of Tokens?

Consider name similarities in Table 6 estimated through equation 2 with homogenous weights (i.e., 1), and variant weights (1, 0.95, 0.90 for α, β and γ respectively). Homogenous weighting scheme estimates same similarity value (i.e., 1) for all pairs of names in Table 6. Is it realistic to say Sim(Ali Daud, Ali Daud) = Sim(A. Daud, A. Daud) = Sim(A. Daud, Ali Daud)? Realistically the probability of two names in record 1 (of Table 6) being to the same person is higher than that of 2’s; and record 2’s probability is higher than that of 3’s. So Sim(Ali Daud, Ali Daud) > Sim(A. Daud, A. Daud), and Sim(A. Daud, A. Daud) > Sim(A. Daud, Ali Daud). To depict our realistic assumption I we employ variant weighting scheme for different types of tokens. It estimates higher similarity value for two names of record 1 than that of those in record 2 and 3 (Table 6, column 5). Same is true for record 2 and 3.

Table 6. name similarities estimated through equation 2.

Sr#	Name 1 (n_i)	Name 2 (n_j)	Sim(n_i, n_j) with Same Weights	Sim(n_i, n_j) with Variant Weights
1	Ali Daud	Ali Daud	1	1
2	A. Daud	A. Daud	1	0.975
3	A. Daud	Ali Daud	1	0.95