ISSN No. 0976-5697

Volume 6, No. 8, November-December 2015

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

On the Use of Fuzzy Clustering in Name Disambiguation

Tasleem Arif
Department of Information Technology
Baba Ghulam Shah Badshah University
Rajouri, India

Abstract: Resolving name ambiguity has become one of the most demanding problems in this era of information overload. This also affects literature management services like digital libraries. It is important to discern ambiguous publications and authors because uncertainty about the real authors of a publication sometimes lead to wrong credits to authors or otherwise. Previous studies have tried to solve this problem by using traditional computational techniques. Soft computing techniques like rough sets, genetic algorithms, fuzzy clustering, etc. promises to be a good option one can look forward to deal with the problems of uncertainty. In this paper, we present the result of our ongoing work for resolving name ambiguity problem in digital citations. We propose a name disambiguation model that uses a mix of hard and fuzzy clustering in a two stage framework. The results of our name disambiguation approach which we obtained on DBLP data are very encouraging and we have been able to achieve very good disambiguation performance in comparison to other baseline methods. Though the results before fuzzy clustering were also very good but after fuzzy clustering the proposed method was able to improve the results. On an average the values of *Precision*, *Recall* and *F1* were *96.35*, *94.01* and *94.72* percent respectively.

Keywords: name disambiguation; ambiguous authors; two stage clustering; digital libraries

I. INTRODUCTION

It is common to find entities with similar or same names and the same applies to virtual world also. In fact the widespread use of virtual platforms like Internet and social media has made the problem more difficult. More and more named entities are appearing on virtual platforms. Named entities form a major part of search queries which search engines serve today. Like the growth of Internet and Internet enabled services the advent of Information Technology has paved the way for proliferation of scientific knowledge [1]. It has been argued [2] that advances in ICT has led to an increase in research productivity, increased level of research collaborations and joint publications between researchers geographically far apart from each other, increase in citations, etc. Thus enabling researchers to collaborate and contribute to the knowledge in their domain of expertise which otherwise would have been difficult. This has also led to accumulation of large amount of bibliographic data in digital libraries like DBLP, CiteSeerX, Microsoft Academic Search, etc. ICT which has made the work of researcher more worthwhile has also compounded the problem for digital libraries by either mixing or splitting the research publications of authors sharing a common name. This is because of the reason that more and more authors with similar names are contributing to scientific knowledge by way of publishing their research work. This is evident from a steep rise in the number of publications in the recent past [3].

In research publications or bibliographies, the name ambiguity problem arises in two different forms, (a) when same name is expressed in different formats and (b), when different authors express their name in similar ways [4]. In first case, the ambiguity arises because of not following a uniform naming pattern by an author. This could happen because of different naming conventions by different journals, conferences, book publishers etc. [5]. A case in point is an author Richard Taylor, Professor Emeritus, Information and Computer Sciences, University of California, Irvine. The publications of Richard Taylor appear under six different name variations: Richard N. Taylor; Taylor, R. N.; R.N. Taylor; Richard Taylor; Taylor, R.; and R. Taylor, even on his homepage[1], leave aside digital libraries. In second case, the ambiguity arises because of multiple authors sharing a common name [5]. This can happen because of limited number of name options that our parents have while choosing a name for us [6]. In DBLP[2] there are nine different Richard Taylor. Along with the Richard Taylor mentioned in the above example, one Richard Taylor is a senior research fellow at Stockholm Research Institute, one with Institute for Information Policy, College of Communications, Pennsylvania State University, one with University of Houston, etc. These problems have long been impeding the efficient information management and retrieval in digital libraries [4]. It requires efficient solutions capable of doing correct attribution and classification of publications of ambiguous authors especially when the information available to deal with such a problem is limited and imprecise in some cases.

The rest of the paper is organized as follows: in second section, we briefly present the related work; in third section, we present our proposed approach for resolving the name ambiguity problem; in fourth section, we present the experimental results, and in the last section, we conclude the paper.

II. BACKGROUND & RELATED WORK

The Referral Web Project of Katz and Martin [7] was an attempt to automatically extract a social network among research of a particular community. During their study they found that it was difficult to differentiate same or similar entities. Though the target audience was limited and the chances of people having same name were quite limited but still they had to look for ways and means to address this problem. Name ambiguity can be seen in a number of fields like digital libraries, web, insurance, etc. It is therefore important to devise mechanisms that could address the name ambiguity. It is not possible to devise a name disambiguation technique that would resolve ambiguity in all areas. Thus the solutions of interest for us here are those techniques that address author name disambiguation.

[1] http://www.ics.uci.edu/~taylor/Publications.htm
[2] http://www.informatik.uni-trier.de/~ley/pers/hs?q=richard+taylor

Efforts for resolving the name ambiguity problem in digital libraries is not a new phenomenon. A number of studies conducted previously have tried to solve the problem. In majority of these cases the techniques proposed so far have broadly been classified under three different headings: supervised learning [8, 9, 10] unsupervised learning [5, 11, 12, 13, 14, 15] and graphic oriented [16, 17].

Supervised techniques try to learn a model based on both positive and negative training examples. Han et al. [18] proposed two name disambiguation models, one based on Bayesian probability, and the other on support vector machines. The technique proposed by Veloso et al. [9] uses a supervised rule based classifier. Peng et al. [10] proposed a model based on Web correlations and authorship correlations using a classifier. These methods try to infer the authors of a publication by using various publication attributes like author(s), title, venue, etc.

Han et al. [5] proposed a K-way spectral clustering based name disambiguation mechanism that uses the same kind of information used by [18]. The method proposed by Masada et al. [12] uses a two-variable mixture model (by adding two variables), an extension of naïve Bayes mixture model. Another unsupervised model proposed by Soler [13] groups publications iteratively based on the similarity between various publication attributes like author(s), e-mail, title, venue, year of publication, keywords etc.

The method proposed by Tan et al. [11] uses a search engine to extract additional information from the Web. On the basis of the information so generated, hierarchical agglomerative clustering (HAC) is used to create clusters of publications. The method proposed by Pereira et al. [14] also obtains additional information from the Web for resolving the author name ambiguity problem. Information is extracted from specific documents on the Web, e.g. CV, by submitting a query to a search engine. The query contains paper title, name of the author and venue. HAC is used to group ambiguous publications which appear on the same Web source. HAC is also used by [15]. The clusters are generated in a bottom-up fashion by first fusing them on the basis of similar co-authors, then title of publication and venue of publication. The process is repeated until no more fusions are possible based on the similarity score.

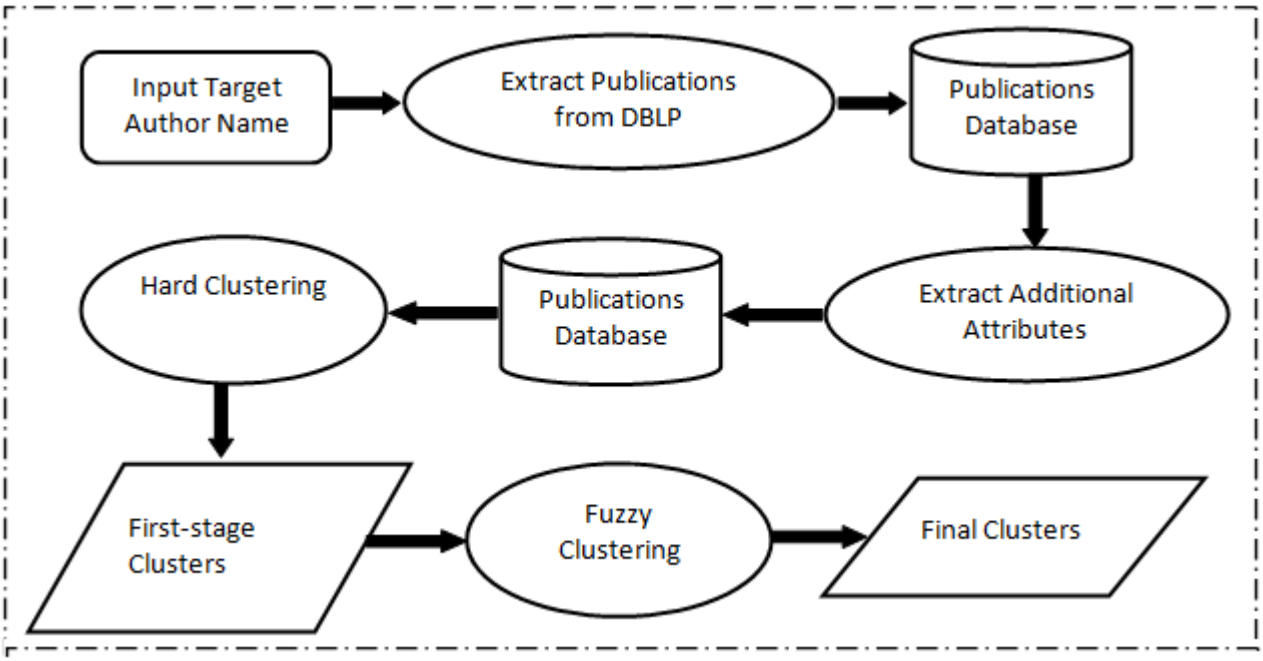

Figure-1: Architecture of the Proposed Name Disambiguation System

The model proposed by Yin et al. [16] applies SVM to weigh different types of linkages used to distinguish authors. In this model what [16] call as DISTINCT, combines two complementary approaches, set resemblance and random walk probability, for measuring similarities between citation records. Another graph theoretic approach, [17] proposed a method called GrapHical framework for name disambiguation (GHOST) using co-authorship information to solve the namesake problem. It first tries to exploit the relationships among publications to construct a graphical model, and solves the namesake problem by serially performing valid path selection, similarity computation, name clustering, and user feedback. GHOST uses only the co-authorship as attribute while excluding all other attributes such as e-mail, publication venue, paper title, and author affiliation, and proposes a novel sophisticated similarity metric to solve the namesake problem.

Unsupervised techniques discussed above use hard clustering mechanism, HAC in majority of the cases. None of these approaches make use of fuzzy clustering. To the best of our knowledge no one till date used fuzzy clustering for name disambiguation in digital libraries. The method proposed by us uses a mixture of hard and fuzzy clustering.

III. PROPOSED NAME DISAMBIGUATION TECHNIQUE

Author name disambiguation can be viewed as a classification problem in which it has to be decided whether the publication under consideration belongs to a particular group or not. Classification methods can broadly follow discriminant analysis or cluster analysis technique. Cluster analysis or clustering (commonly known term) is used in those situations where little or no information is available about group structure prior to the classification [19].

Traditional clustering methods have been used for author name disambiguation in a number of different ways [20]. In the proposed approach we use a mix of hard and fuzzy clustering in a two stage clustering framework. In the first stage we use hard clustering framework and in the second we use fuzzy clustering framework. Figure-1 shows the architecture of the proposed system. The bibliographic data for an author name is extracted

from DBLP using the methodology shown in Figure-2. After extraction this data is supplemented with additional publication attributes obtained in a resource bound manner [21] from WWW using a search engine. We do not go into the details of the extraction of the additional publication features.

In first stage we use the clustering process and similarity measures used in [6]. In second stage, we compute the distance between all the available attributes of a publication with those of the other to combine these distances into a similarity score. This similarity score is used to calculate the value of membership function used for fuzzy clustering step.

Fuzzy or soft clustering allows data elements to belong to more than one cluster simultaneously, and be associated with each cluster with certain membership levels. The degree or grade of membership which can be any value in the range [0, 1] indicates the strength of the association between that data element and a particular cluster. Soft clustering is a process of assigning these membership values, and then using these membership values to assign data elements to one or more clusters. Soft clustering has proved to be beneficial in dealing with uncertainty. There may be certain cases where agglomerative clustering used in first stage may have a good number of clusters with only one citation record. In such a case we use fuzzy clustering to find the relative similarity between clusters having a single publication and the rest by calculating the value of membership function ($\boldsymbol{\mu}$) by using Equation (1) as follows:

$$\mu_{i,j}(cr_i, C_j) = \cos(cr_i, C_j)^m \Big/ \sum_{r=1}^{k} \cos(cr_i, C_r)^m \quad (1)$$

where cr_i is the i^{th} publication in a singleton cluster and C_j is the j^{th} cluster ($i \neq j$), respectively, and m is the fuzzy factor.

The parameter m determines the "softness" of the clustering solution. If $m=0$, the degree of membership of a publication with all the remaining clusters is same and when m approaches ∞, the clustering becomes hard clustering [22]. In general, the softness of the clustering solution is inversely proportional to fuzzy factor m. In our case, we merge a singleton cluster with any other cluster only if the value of fuzzy membership function is above a threshold.

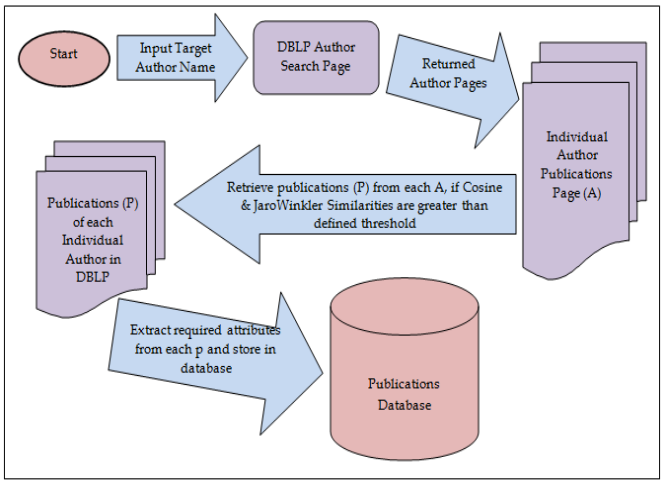

Figure-2: Publication Data Extraction from DBLP

IV. Experimental Results

In order to test the efficiency of the proposed approach publications metadata of ten ambiguous authors (author names) was extracted from publications of indexed by DBLP. The statistics of the dataset used are shown in Table-I. Here, *Pubs* refer to the number of publication records retrieved from DBLP for the author name listed in a particular record, *A-Authors* to the number of real authors, *P-Authors-S1* to the number of authors predicted by the proposed approach after hard clustering stage i.e. first stage and *P-Authors-S2* to the number of authors predicted after fuzzy clustering stage i.e. second stage. This included 1527 publications of 137 real life authors.

Table I. Author Dataset

Author	***Pubs***	***A-Authors***	***P-Authors-S1***	***P-Authors-S2***
David Jensen	82	5	5	5
Charles Smith	40	15	16	16
Michael Wagner	232	20	25	19
Robert Moore	91	12	18	18
Hui Fang	156	21	25	23
Jie Tang	227	12	13	12
Richard Taylor	186	19	24	20
William Cohen	190	5	7	7
Joseph Miller	28	4	6	4
Gang Wu	295	24	40	37
Total	***1527***	***137***	***179***	***161***

The performance of the proposed disambiguation approach used in this study has been shown in terms of percentage Precision, Recall and F1 scores in Table-II and Table-III. These metrics are used in the same way as they have been used in [6] for the same purpose. Table-II presents the name disambiguation results in terms of the considered metrics i.e. precision, recall and F1 before the application of fuzzy clustering. Thus the results presented in Table-II are the results obtained after first stage of clustering.

Table II. Name Disambiguations Results After First Stgae

Author	***Predicted Authors***	***Precision***	***Recall***	***F1***
David Jensen	5	100	100	100
Charles Smith	16	97.50	100	98.73
Michael Wagner	25	96.70	90.71	93.61
Robert Moore	18	91.21	100	95.40
Hui Fang	25	97.39	98.03	97.70
Jie Tang	13	97.80	100	98.89
Richard Taylor	24	89.89	95.24	92.49
William Cohen	7	98.95	100	99.47
Joseph Miller	6	89.29	100	94.34
Gang Wu	40	87.50	54.65	67.28
Average	***179***	***94.63***	***93.86***	***93.80***

The values of Precision, Recall and F1in Table-II are 94.63, 93.86 and 93.80, respectively. These results are those obtained

after hard clustering stage. It can be observed that there are good numbers of authors whose publications are fragmented and have been grouped in different groups.

Table-III presents the values of the above metrics for all the authors and the percentage change in the values of these metrics over their respective values obtained in the first stage i.e. hard clustering stage. Although these values may not represent any significant change but whatever they have been able to achieve is quite meaningful for the disambiguation process. In case of Michael Wagner, the fuzzy clustering step has been able to improve the values of precision, recall and F1 by a margin of 3.3, 4.79 and 4.09 percent respectively. On an average the improvements in precision, recall and F1 for all the ten authors is 1.72, 0.15 and 0.92 percent. It may seem that the improvements are negligible but it is hard to improve the performance when the results are already more than ninety percent.

Table III. Name Disambiguations Results After Fuzzy Clustering

Author	***Predicted Authors***	***Precision***	***Recall***	***F1***
David Jensen	5	100	100	100
Charles Smith	16	97.50	100	98.73
Michael Wagner	19	100	95.50	97.70
Robert Moore	18	91.21	100	95.40
Hui Fang	23	98.04	97.40	97.72
Jie Tang	12	98.23	99.55	98.89
Richard Taylor	20	90.45	93.60	92.00
William Cohen	7	98.95	100	99.47
Joseph Miller	4	100	100	100
Gang Wu	37	89.09	54.04	67.28
Average	***161***	***96.35***	***94.01***	***94.72***

For comparison of name disambiguation results this study considered HAC [11] which uses agglomerative clustering and the publications metadata is augmented using search engine results in a similar fashion that we used in the first stage. The comparison of the results obtained through the proposed approach with the base line method taken from [23] on all three metrics listed above is shown in Table-IV.

Table IV. Comparison with HAC

Author	***HAC***			***FMC***		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
David Jensen	85.85	94.88	90.14	100	100	100
Charles Smith	30.00	100	46.15	97.50	100	98.73
Michael Wagner	18.35	60.26	28.13	100	95.50	97.70
Robert Moore	86.90	93.10	89.89	91.21	100	95.40
Hui Fang	100	100	100	98.04	97.40	97.72
Jie Tang	100	100	100	98.23	99.55	98.89
Richard Taylor	80.17	99.93	88.97	90.45	93.60	92.00
William Cohen	81.53	97.98	89.00	98.95	100	99.47
Joseph Miller	54.55	54.55	54.55	100	100	100
Gang Wu	97.54	97.54	97.54	89.09	54.04	67.28
Average	***73.49***	***89.82***	***78.44***	***96.35***	***94.01***	***94.72***

The values of precision, recall and F1 obtained through the proposed approach were 96.35, 94.01 and 94.72 percent, respectively which is huge improvement over the baseline method that we used to compare our proposed approach. In case of majority of the authors under consideration, the values of all the three metric were more than 90 percent. In case of Gang Wu, the low value of recall was instrumental in bringing down the value of F1 to 67.28. The low value of recall in this case can be attributed to large number of false-negative cases as more than one Gang Wu published in a similar venue. The proposed approach has been able to improve the values of precision, recall and F1 by 22.86, 4.19 and 16.28 percent respectively over HAC.

V. CONCLUSIONS

Author name disambiguation is an important area of research with increasing number of efforts dedicated to address it. In case of digital citations and digital libraries resolving name ambiguity has become important in view of increasing number of publications and widespread usage of digital libraries among the researchers. In this paper we proposed a hybrid clustering mechanism by employing hard clustering in first stage and soft clustering in the second. Experimental results conducted on DBLP dataset are very encouraging as the proposed approach has been able to achieve F1 score of 94.72 percent. By using soft clustering we have been able to deal with the split citation problem to a good extent. In some cases where F1 score were below expectations is due to the fact that more than one authors published with same journals or conferences which lead to distinct clusters being merged based on venue information. We are also of the view that with ever increasing number of publications with similar authors venue information may not prove to be a good feature for disambiguation purposes.

VI. REFERENCES

[1] H.W. Chang and M.H. Huang, “Cohesive subgroups in the international collaboration network in astronomy and astrophysics,” Scientometrics, 2014, 101(3), pp. 1587-1607.

[2] D. Zhao and A. Strotmann, “The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis,” Journal of the Association for Information Science and Technology, 2014, 65(5), pp. 995–1006.

[3] L. Tang, and J.P. Walsh, “Bibliometric fingerprints: name disambiguation based on approximate structure and equivalence of cognitive maps,” Scientometrics, 2010, 84(3), pp. 763-784.

[4] D. Shin, T. Kim, J. Choi, and J. Kim, “Author name disambiguation using a graph model with node splitting and merging based on bibliographic information,” Scientometrics, 2014, 100(1), pp. 15-50.

[5] H. Han, H. Zha, and C.L. Giles, “Name disambiguation in author citations using a K-way spectral clustering method,” In Proceedings of Joint Conference on Digital Libraries’2005, pp. 334 – 343.

[6] T. Arif, R. Ali, and M. Asger, “Author name disambiguation using vector space model and hybrid similarity measures,” In Proceedings of 7th International

Conference on Contemporary Computing-IC3'2014, Noida, India: IEEE, pp. 135-140.

[7] H. Kautz, B. Selman, and M. Shah, "The hidden web.," American Association for Artificial Intelligence magazine, 1997, 18(2), pp. 27–35.

[8] H. Han, H. Zha, and C.L. Giles, "A model-based K-means algorithm for name disambiguation," Proceedings of 2nd International Semantic Web Conference, USA, 2003.

[9] A. Veloso, A. A. Ferreira, M. A. Gonçalves, H.F.A. Laender, and W. Meira Jr., "Cost-effective on-demand Associative Author Name Disambiguation," Information Processing and Management, 48(4), 2012, pp. 680– 697.

[10] H. Peng, C. Lu, W. Hsu, and J. Ho, "Disambiguating authors in citations on the web and authorship correlations," Expert Systems with Applications, 39(12), 2012, pp. 10521-10532.

[11] Y.F. Tan, M. Kan, and D. Lee, "Search engine driven author disambiguation," Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL '06), 2006, pp. 314-315.

[12] T. Masada, A. Takasu, and J. Adachi, "Citation data clustering for author name disambiguation," In Proceedings of 2nd International Conference on Scalable Information Systems, 2007.

[13] J. Soler, "Separating the articles of authors with the same name," Scientometrics, 72(2), 2007, pp. 281-290.

[14] D.A. Pereira, B. Ribeiro-Neto, N. Ziviani, A.H. Laender, M.A. Gonçalves and A.A. Ferreira, "Using web information for author name disambiguation," In Proceedings of 9th ACM/IEEE-CS Joint Conference on Digital Libraries'2009, ACM.

[15] R.G. Cota, A.A. Ferreira, C. Nascimento, M.A. Gonçalves and A.H.F. Laender, "An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations," Journal of the American Society for Information Science and Technology, 61(9), 2010, pp. 1853–1870.

[16] X. Yin, J. Han, and P.S. Yu, "Object distinction: Distinguishing objects with identical names," In Proceedings of IEEE International Conference on Data Engineering, 2007, pp. 1242-1246.

[17] X. Fan, J. Wang, X. Pu, L. Zhou and B. LV, "On graph-based name disambiguation," ACM Journal of Data and Engineering Quality, 2(2), 2011, pp. 10.

[18] H. Han, L. Giles, H. Zha, C. Li and K. Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations." In Proceedings of Joint Conference on Digital Libraries'2004, pp. 296 – 305.

[19] T. Naes and B-H. Mevik, "The flexibility of clusters illustrated by examples," Journal of Chemometrics, 13(4), 1999, pp. 435-444.

[20] T. Arif, M. Asger and R. Ali, "Author name disambiguation using two stage clustering," INROADS-An International Journal of Jaipur National University (Special Issue), ISSN: 2277-4904, 3(1), 2014, pp. 340-345.

[21] P. Kanani, A. McCallum and C. Pal, "Improving author coreference by resource-bounded information gathering from the web," Proceedings of 20th International Joint Conference on Artificial Intelligence-IJCAI, Hyderabad, India, 2007, pp. 429-434.

[22] Y. Zhao and G. Karypis, "Soft clustering criterion functions for partitional document clustering: a summary of results" Proceedings of 13th ACM International Conference on Information and Knowledge Management, New York, USA, 2004, pp. 246-247

[23] J. Tang, A.C.M. Fong, B. Wang and J. Zhang, "A unified probabilistic framework for name disambiguation in digital library," IEEE Transactions on Knowledge and Data Engineering, 24(6), 2012, pp. 975-987.

