International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217358

Importance of Name Disambiguation in Scientific Databases

Tasleem Arif [1,2], Majid Bashir Malik[3]

[1]Department of Information Technology, BGSB University, Rajouri, J&K, India

[2]Department of Computer Science, Shaqra University, Kingdom of Saudi Arabia

[3]Department of Computer Sciences, BGSB University, Rajouri, J&K, India

Article Info
Volume 3, Issue 7
Page Number: 487-492

Publication Issue :
September-October-2018

Article History
Accepted : 02 Sep 2018
Published : 20 Sep 2018

ABSTRACT

Ambiguity in digital citation databases is a major bottleneck in attribution of proper credit to authors and thus hampers the process of profiling authors in true sense. It is quite common for academics and researchers to share common or similar names and the recent surge of digital citation records has amplified the problem exponentially. Realizing the prowess of information and communication technologies and the ease with which the information can be stored, managed and shared online, traditional publishers and databases have joined the bandwagon and embarked on the journey of digitizing their records. In the absence of an effective mechanism, it becomes extremely difficult for a computer to discriminate between similar entities and more so in case of our names. This paper highlights some of the major advantages and drawbacks of prominent categories of solutions by supporting the inferences with relevant backups, wherever required.

Keywords : Name Disambiguation, Digital Libraries, Scientific Databases, Ambiguous Author References

I. INTRODUCTION

In any society there has always been a sort of scarcity of name options which one can choose from to give name, which is a basic identification, to our newborn. This not so seemingly a matter of concern for humans poses a major challenge in the world being increasingly dominated by computers. Humans have always been better in differentiating similar things in a much better way which at times is based on background knowledge. This becomes quite a herculean task for automated computer based systems. Therefore some sort of training or other measure needs to be adopted to effectively discriminate named entities.

Due to diverse factors the academic activities have received considerable impetus in last few decades. This has led to exponential rise in publication of books, articles and other academic or research material. Publication of research findings in the form of journal articles, conference papers, book chapters, etc. is a major activity of the people working in higher education institutions and R&D institutions worldwide. Since these publications are being increasingly stored in the form of digital

citations by various scientific databases [4], differentiating between similar named authors has become a cause of concern for various services in this domain. This has led to proposal and development of various name disambiguation techniques based on varying consideration, with equally varying efficiencies [1, 2].

Despite numerous efforts, some of which are currently being employed by digital citation databases, the problem is still looming large and it is not very rare to find papers wrongly listed under the profiles of same or similar named authors. The ambiguity in author names results in two scenarios, a) where the publications of an individual are listed under more than one author names, which is referred to as 'split citation problem' and b) where the publications of two or more individuals are listed under a single author, which is referred to as 'mix citation problem'. Here an individual means a real life singular entity i.e. human and author means an entity identified by the citations database.

In this paper an attempt has been made to identify prominent reasons behind the existence of ambiguity either in the form of split citation or mixed citation problem and a summary of prominent and path-breaking techniques in the concerned domain is presented. The main aim of this article is to help new researchers in this area to help them bootstrap their efforts and provide them with a gist of representative techniques.

II. ATTRIBUTES USED FOR NAME DISAMBIGUATION

Studies available in the public domain aimed at author name disambiguation have listed various attributes of a publication which can be employed in diversified combinations for obtaining desirable results. Traditional attributes forming the metadata of a publication are readily available however, obtaining additional or privileged attributes and information is a resource intensive task. A summary of these attributes, their expected effect and availability is provided in [3]. A total of eight different publications' attributes viz. title (of publication), author(s), email ID(s), affiliation(s), venue, year-of-publication, references and contents have been listed in [3]. Though some of these attributes like title, authors, venue, year-of-publication and references can be obtained easily, others like, affiliations, email IDs, and contents may not be available for all of the publications of an individual without hassles. In addition to their availability the efficiency of all of these attributes is a major consideration. It is quite vivid that more the number of attributes used by a technique for disambiguation purposes, the more it's time and space complexity will be. Thus researchers working in this domain have to choose with extreme caution the type and number of attributes to make their proposed techniques relevant for large scale practical usage [3].

III. MAJOR NAME DISAMBIGUATION TECHNIQUES, THEIR ADVANTAGES AND DRAWBACKS

Digital scientific databases list millions of citation records which mostly include a list of authors, the title of the publication and some other attributes. These attributes are not disjoint in true sense as common terms like name(s) of author(s), publication venue, year, etc. may be common between a huge number of these digital citation records or publications for simplicity. Due to this inherent property ambiguity is bound to happen and it can lead to any of the two cases listed above. Owing to this ambiguity in publications has received significantly high attention from the research community. Thus a number of advanced techniques to solve this problem have been proposed in the literature as evidenced by two promising summaries of work conducted on author name disambiguation

[2, 17]. Though techniques proposed so far have listed the existing techniques under various headings, these techniques have mainly been classified under two categories based on their dependency on the training set. If the technique depends upon the training set, it is classified as Supervised, other Unsupervised.

Machine learning plays a crucial role in case of supervised techniques [8, 9, 10] where the disambiguation model is trained beforehand and once the requisite training is imparted the trained model can be put to use to disambiguate ambiguous authors. To achieve better efficiency and efficacy both positive and negative labeled information has used to train the model in some of the proposed studies. Such techniques assign the references to their authors by employing a supervised machine learning technique. In other words a set of references to authors with their attributes, the training data D is provided as input, which is positively labeled i.e. references for which the correct authorship is known. Each example is composed of a set X of n attributes $\{X_1, X_2, \ldots, X_n\}$ along with a special variable called the author. This author variable 'a' draws its value from a discrete set of labels $\{a_1, a_2, \ldots, a_n\}$, where each label uniquely identifies an actual author. The examples used in the training generate a disambiguation function that maps the attributes in the training examples to the correct author. On the other hand the test set (T) for the ambiguity resolution task consists of a set of references for which the attributes are known while the correct author is unknown. In this process the job of disambiguation function is to correctly derive the correct mapping between the attributes and the authors based on the references in the test set.

It has been observed that supervised learning based author name disambiguation techniques produce excellent results when they are subjected to a large number of examples of citations for each author [2]. In addition these methods have the advantage of performing the disambiguation in incremental fashion, wherein, if the existing collection of references has been disambiguated by either using automatic or manual means, new citations can be efficiently disambiguated by employing the trained model. Despite their efficiency as evidenced by their practical use (as reported in the relevant literature) the acquisition of collection of labeled examples for each author becomes a herculean task in view of exponential rise in the number of authors producing ever increasing publications [5]. Application of such methods in case of digital libraries may not result in expected results owing to their dynamic nature. In such cases keeping pace with the changes in the publication environment may not be possible for such extremely large systems maintaining data about diverse authors, subjects, areas, etc. Moreover, training the model for unavoidable change in author interests over time and incorporating each such training model may not be feasible on one hand and scalable on the other.

However most of the unsupervised methods use a clustering algorithm to group publications on the basis of a distance function [19]. Detailed discussions about these methods can be found in [1, 2, 5, 17, 18, 19, 20]. Techniques in this category of disambiguation try to directly assign publications to authors by optimizing the fit between a set of references to an author and some mathematical model used to represent that author. Clustering techniques use a probabilistic framework to determine the author in an iterative way to fit the model of the authors. The process of resolving ambiguity works in stages or iterations where each step or iteration use one or a combination of attributes to fine tune the results for the next iteration. This process continues until a stop condition is reached, for instance, after a number of iterations or all the publication attributes under consideration have been used to derive the disambiguated results. Expectation-Maximization

(EM) [6] and Gibbs Sampling [7] are commonly used in such cases.

These methods have the advantage of being able to directly assign a new addition to the already disambiguated publications for an author however their dependence on some sort of special information like correct number of authors which may not be available in advance, thus limiting their potential use where it may not be either possible or practical to guess the exact number of distinct authors. Another drawback of clustering based name disambiguation techniques is that they tend to be slower in action as compared to their supervised counterparts because of the number of iterations through which each new reference will have to go through.

The techniques classified above use diverse combinations of basic information or traditional publication attributes like name author(s), title and publication venue, however, the rise in academic and scientific productivity has made it almost infeasible to derive meaningful results by just employing these basic attributes. In the quest to improve the performance of disambiguation task some studies [5, 11, 12, 13, 14, 15, 16] have successfully proposed the use additional attributes or implicit information from varying sources on the Web for effective disambiguation. Torvik et. al. [11] proposed a probabilistic model for author name disambiguation in Medline. This model uses affiliation of author(s) in addition to traditionally used publication attributes in Medline for calculation of pairwise similarity score between two ambiguous papers. The disambiguation model proposed by Kanani et al. [12] uses additional information from the Web and view the disambiguation task as a graph partitioning problem. The additional information is obtained in a resource bound manner to prevent information overload and the information obtained from search results is incorporated as additional feature or as an additional node in the graph. Additional information obtained from the Web improves the efficiency of disambiguation considerably [13]. Yang et al., [13] uses Web correlation in addition to Topic correlation, where similarity between a pair of ambiguous publications is estimated from their correlation on the Web. Implicit information about co-authors of conflicting publications is obtained from the Web [14] and explicit information from Web pages and curricula vitae of target authors is obtained and used for disambiguation in [15].

The model proposed by D'Angelo et al. [16] uses additional information about affiliation and research areas of authors to improve the performance of disambiguation in a specific domain. The additional information used in [5] is e-mail ID and affiliation of authors. These additional attributes were obtained from PDF of publications where these were available however in case PDF is not available affiliation information was obtained from select sources on the Web, if available.

IV. CONCLUSIONS

All the major name disambiguation techniques including supervised and unsupervised or those using additional information have their advantages and drawbacks. The supervised techniques are good at disambiguation provided the situation remains stagnant whereas despite being slow the unsupervised techniques have the potential to disambiguate new additions in a much better fashion. Despite additional attributes or publication information having a profound impact on the disambiguation performance of author name disambiguation, the use of additional attributes faces various inherent hindrances. The primary concern is their availability [2, 3]. In most of the digital libraries traditional attributes like author(s), title, publication venue, year-of-publication and a list of reference used by a particular publication form the

metadata about publications they store. This means that techniques proposing the use of additional attributes or information, as listed in [3], have to devise some effective and efficient mechanism to gather or extract the required additional information. For instance, Springer database includes the affiliation(s) of author(s) but the viewer has to expand that information by clicking on the provided hyperlinks. For an automated process clicking on the right hyperlink, for extraction of affiliation(s) information of author(s) adds a lot to the complexity of the process. The extraction/gathering of other additional publication attributes like keywords, etc. may be more cumbersome in some cases. Thus before deciding on the use of this additional or privileged information one has to strike a balance between efficiency and accuracy. In practical situations, such as disambiguating publications digital derived from a digital library, additional attributes or publication information can be used on an adhoc basis without employing too much efforts and resources for gaining access to this additional information.

V. REFERENCES

[1]. Hussain, I., and Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. The Knowledge Engineering Review, 32, E22. doi:10.1017/S0269888917000182

[2]. Ferreira, A.A., Gonçalves, G.A., and Laender, H.F.A. (2012). A brief survey of automatic methods for author name disambiguation. ACM SIGMOD Record, 41(2), pp: 15-26.

[3]. Arif, T., Ali, R., and Asger, M. (2015). A multistage hierarchical method for author name disambiguation, International Journal of Information Processing, 9(3), pp: 92-105.

[4]. Coscia, M., Giannotti, F., Pensa., R. (2009). Social Network Analysis as Knowledge Discovery process: A case study on Digital Bibliography. Proceedings of the Advances in Social Network Analysis and Mining, 2009, pp: 279-283.

[5]. Arif, T., Ali, R., and Asger, M. (2014). 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.

[6]. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B, 39(1), pp:1–38, 1977.

[7]. Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. The National Academy of Sciences, 101(1), pp: 5228–5235, 2004.

[8]. Han, H., Giles, C. L., Zha, H., Li, C. and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In Proceedings of 2004 JCDL, pp: 296–305, 2004.

[9]. Veloso, A., Ferreira, A. A., Gonçalves, M. A., Laender, A. H. F. and Meira Jr., W. (2012). Cost-effective on-demand associative author name disambiguation. Information Processing and Management, 48(4), pp: 680– 697, 2012.

[10]. Ferreira, A. A., Veloso, A., Gonçalves, M. A., and Laender, A. H. F. (2010). Effective self-training author name disambiguation in scholarly digital libraries. In Proceedings of 2010 JCDL, pp: 39–48, 2010.

[11]. Torvik, V.I., Weeber, M., Swanson, D.R., and Smalheiser, N.R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research articles. Journal of the American Society for Information Science and Technology, 56(2), pp: 140–158.

[12]. Kanani, P., McCallum, A., and Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the web.

In Proceedings of 20th International Joint Conference on Artificial Intelligence-IJCAI, Hyderabad, India, pp: 429-434.

[13]. Yang, K.-H., Peng, H.-T., Jiang, J.-Y., Lee, H.-M., and Ho, J.-M. (2008). Author name disambiguation for citations using topic and web correlation. In B. Christensen-Dalsgaard, D. Castelli, B.A. Jurik, & J. Lippincott (Eds.), Research and advanced technology for digital libraries (pp. 185–196). Berlin Heidelberg: Springer.

[14]. Kang, I.-S., Na, S.-H., Lee, S., Jung, H., Kim, P., Sung,W.-K., and Lee, J.-H. (2009). On co-authorship for author disambiguation. Information Processing & Management, 45(1), pp: 84–97.

[15]. Pereira, D.A., Ribeiro-Neto, B., Ziviani, N., Laender, A.H., Gonçalves, M.A., and Ferreira, A.A. (2009). Using web information for author name disambiguation. Paper presented at the Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM.

[16]. D'Angelo, C.A., Giuffrida, C., and Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. Journal of the American Society for Information Science and Technology, 62(2), pp: 257–269.

[17]. Smalheiser, N.R., and Torvik, V.I. (2009). Author name disambiguation. Annual Review of Information Science and Technology, 43(1), pp: 1–43.

[18]. Tang, J., Fong, A.C.M., Wang, B., and Zhang, J. (2012). "A Unified Probabilistic Framework for Name Disambiguation in Digital Library." IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 6, June 2012, pages: 975-987.

[19]. Liu, Y., Li, W., Huang, Z. and Fang, Q. (2014). A fast method based on multiple clustering for name disambiguation in bibliographic citations. Journal of the Association for Information Science and Technology. DOI: 10.1002/asi.23183

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

Cite this article as :

Tasleem Arif, Majid Bashir Malik, "Importance of Name Disambiguation in Scientific Databases", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 3 Issue 7, pp. 487-502, September-October 2018. Available at
doi : https://doi.org/10.32628/CSEIT217358
Journal URL : https://ijsrcseit.com/CSEIT217358