Review on A Graph Based Multi-level Clustering for Author Name Disambiguition

[1]Khushbu G. Panpaliya, [2]Prof. M. S. Burange, [3]Prof. P. D. Soni
[1]Student, [2]Professor, [3]Professor
[1]Department of Computer Science & engineering,
[1]P. R. Pote College of Engineering, Amravati, Maharashtra, India

Abstract— For any literature, the basic thing is to identify individual(s) who wrote it or to identify all the research work related to individual author. Attribution would seem to be simple process but yet it represents a major unsolved problem for data mining. Researchers use generally scholarly digital libraries for their relevant research work as digital libraries have ubiquitous availability of scholar articles and publications. When someone tries to access an article using author name result produced may not meet user's expectation due to name ambiguity. Ambiguous names often lead to confusion and mistakes in identification of author's research work. And this leads to author name disambiguation process. Another case in which author name ambiguity can be seen is the retrieving the papers of an author who have used distinct name variations in different articles. Furthermore, name disambiguation for web papers can be even more challengeable with increasing mentioning of ambiguous names.

IndexTerms— Disambiguation, Clustering, Multi-level Clustering, Supervised learning, Unsupervised learning

I. INTRODUCTION

Now-a-days, researchers generally use different digital libraries such as Google Scholar, Microsoft Academic Search, etc. for their relevant research work. Whenever, the most commonly executed query of digital libraries i.e. author name search is executed, the result produced is of impaired quality or may not satisfy user's expectation due to author name ambiguity. Ambiguous names often lead to confusion and mistakes in identifying records related to author's research work. In order to improve quality of research work, author name disambiguation is performed.

Author name disambiguation separates the cases of ambiguous names referring to distinct authors and merging cases of variant names referring to same individual across all authors and papers. Author name disambiguation comprise of four distinct challenges: First, an author may uses multiple names for different publications, this includes, orthographic and spelling variant or spelling errors in author name, name change of author over time due to marriage for female authors or due to religious conversion or gender re-assignment. Second, several authors with same name, in fact, common names may comprise several thousand authors. Third, necessary attributes of author entity may incomplete or entirely unavailable due to a reason some publishers may not recorded author's first name, their geographical locations or identifying information such as their degrees or their positions, etc. Last, an increasing percentage of scholarly articles not only multi-authored but also multi-disciplinary and multi-institutional efforts. In such cases, disambiguating some authors does not necessarily help assign the remaining authors [8].

So, author name disambiguation is not trivial and straightforward task. In order to resolve ambiguity algorithmic approaches can be used. Algorithmic methods are challenging for two reasons: First, they have to rely on metadata and metadata for large scale databases is often sparse especially for old applications. Second, disambiguation algorithms may draw false conclusions when faced with incomplete metadata. This issue can be present in any case where an individual attributes are not consistent over time.

II. EXISTING SYSTEM & THEIR LIMITATIONS

Numerous workshops have been called for effective disambiguation methods. For example Author id meeting [1]. Many individual publishers and researchers have set up their own internal disambiguation efforts on a massive scale. These activities show importance of author name disambiguation in data mining. Disambiguation is needed to create link from digital libraries to online sources. Some of the different approaches proposed are as follows:

Registry of unique author identifiers

In February 5, 2007 a meeting was called by Crossref to determine whether registry of author can solve ambiguity problem in data mining. UAI_Sys which so far implemented a pilot project, in that, each author submits the list of its pre-existing publications when joins the system, it would allow one to assign the many articles. Author would enter their own metadata and set their own passwords and would be assigned with 16 digit unique id number. Then author can use this number for all their publications. It is assumed that authors will agree to remember passwords and update the metadata periodically [1].

Although the scheme has conceptual simplicity and it is technically feasible, it fails to take into account the realities of human behavior. Authors not only have to cooperate actively but also they have to update metadata periodically. For this vast majority of authors have to participate even those authors who wrote only single article. It is likely that registry will garner universal support by authors who do not receive any reward for participating. This scheme also fails to take into account the tenuous nature of web-based resources and their funding [2].

Manual disambiguation

Most cases of author name disambiguation refers to manual curation. For example, Mathematical reviews have disambiguated over 2 million publications manually since 1940 [3]. Several initiatives make use of combination of automatic or author supplied or community supplied input. For example, DBLife (DeRose et al., 2007) extracts author information from within a defined database research community, and displays it in a standardized format that is subject to manual correction. Several web based services allows authors to register and create profiles to link their papers. For example, community of science has almost 480,000 profiles and has about 15000 registered authors. Nevertheless manual disambiguation is hard and uncertain process even on small case, it is infeasible for common names.

Machine learning approach

Research approaches are machine learning approaches. These automatic machine learning approaches for author name disambiguation are categorized as supervised, semi-supervised and unsupervised. Supervised approach of disambiguation automatically learns multi-category classifiers for each ambiguous author name from annotated data to predict corresponding author entity for each paper. To train such classifiers, information such as titles, co-authors, and venues can be found directly from citation records, additional information such as abstract and affiliations can be extracted from the content of each paper. Classifiers can be trained through Hybrid Naïve Bayes, Support vector machines or other methods [4]. The technique proposed by Veloso et al. [5] uses a supervised classifier. Peng et al. [6] proposed a model based on web correlations using a classifier.The method proposed by Masada et al. [9] uses a two-variable mixture model (by adding two variables), an extension of Naïve Bayes mixture model. So, the constraint to use supervised learning approach is that it requires large amount of trained corpus. To obtain trained corpus is time consuming and also money intensive and labor intensive. Hence, these approaches are not practical for large scale of data.

Another machine learning approach for author name disambiguation is semi-supervised disambiguation. This approach requires only seed amount of trained classifiers. By using this small amount of trained corpus it performs disambiguation. It possess all advantages and disadvantages of supervised learning approach

In absence of annotated data and trained classifiers, unsupervised approach of disambiguation is used. Unsupervised approach manages to find matching between papers and real author entities using clustering algorithms or topic models. Clustering algorithms take pairwise similarity functions to group papers into clusters as different entities. Similarity functions can be predefined based on existing features or domain knowledge or can be learned from supervised learning algorithms. A graph theoretic approach, [10] 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. Another approach proposed by Tasleem Sharif in [11] uses fuzzy clustering for author name disambiguation.

III. PROPOSED SYSTEM

Author name disambiguation can be viewed as a classification problem in which input is mapped with some discrete values on the basis of certain decisions i.e. input author name is mapped with related publications on the basis of certain decisions that belongs to group or not and result is produced. The proposed system is the unsupervised machine learning approach for author name disambiguation. Following figure Fig-1 gives 6-step architecture of proposed system.

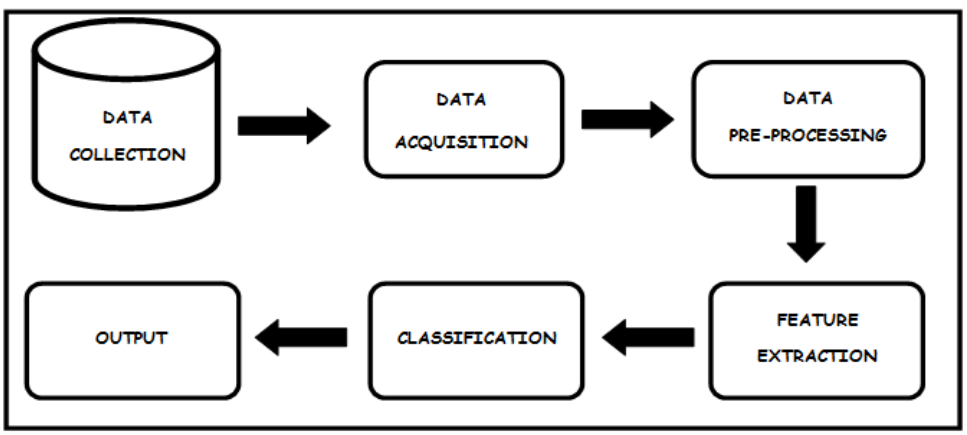

Fig-1: Architecture of the Proposed System

First step of proposed system will be data collection. In this database will be constructed which includes collection of records of different authors with their publications. Metadata relevant to authors and publications will be as fields of database. The second step will be the data acquisition. This phase acts as input phase of the system. In this the most common query related to digital libraries, that is, author name search is executed. Third step will be the data pre-processing. After executing a query result is produced with ambiguity and we process that result for disambiguation process, this is our pre-processing. Fourth is the feature extraction. In this features or attributes related to entities such as title, publication year, author name, co-author name, etc. are extracted from source data and this can helps in simplifying clustering process. The fifth step is the classification in which we used Sequential k-Means clustering algorithm for clustering relevant data and mapped to produce result. And sixth and the last step is the output phase in which ambiguity free result is obtained.

IV. APPLICATIONS OF PROPOSED SYSTEM

1. Proposed system may be used in development of semantic web.
2. The system may used as wrapper on top of many digital libraries.
3. The system may used to find out all research work related to a author.
4. The system may used for large scale databases.

REFERENCES

[1] http://www.crossref.org/CrossTech/2007/02/crossref_author_id_meeting.html
[2] Neil R. Smalheiser and Vetle I. Torvik., "Author Name Disambiguation" in: Annual Review of Information Science and Technology , Inc: http://books.infotoday.com/asist/#arist, Vol. 43, 2009
[3] http://www.ams.org/mr-database/mr-authors.html
[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p/p 109.
[5] 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, p/p. 680– 697.
[6] 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, p/p. 10521-10532.
[7] 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, p/p. 296 – 305.
[8] Simeng Sun, Hui Zhang, Ning Li, Yong Chen. "Name Disambiguation for Chinese scientific authors with multi-level clustering", In (EUC) ISSN No:1709-8728 p/p: 176-182,2017
[9] T. Masada, A. Takasu, and J. Adachi, "Citation data clustering for autor name disambiguation", In Proceedings of 2nd International Conference on Scalable Information Systems, 2007.
[10] 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.
[11] Tasleem Sharif, "On the Use of Fuzzy Clustering in Name Disambiguation", In Proceedings of international journal of advanced research in computer science, ISSN No: 0976-5697, p/p. 53-57, 2015.