Implementation of A Graph Based Multi-level Clustering for Author Name Disambiguiation

Ku. Khushbu Panpaliya[1], Prof. Mayur S. Burange[2], Prof. Pravin D. Soni[3]
P. R. Pote College of engineering and management, Amravati

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.*

Keywords- Multi-level clustering; Disambiguation; Discipline tree; 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.

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. 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. The proposed system is developed to perform author name disambiguation for digital libraries which helps to get quality result and user satisfied result of author name search query. For this from earlier work done it is analyzed that the primary need is proper database having large amount of publications with authors sharing same name or ambiguous authors publications. Following Figure shows the data flow diagram of proposed system.

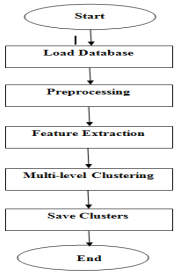

Figure: Data Flow Diagram of Proposed System

Step-wise Workflow of Proposed System:

Step 1: Load Database
For author name disambiguation several databases are available, so according to that database, it could have following attributes-
Author, co-author, title, journal, archive Prefix, eprint, primary class, year, month, etc. The proposed system uses AND (Author Name Disambiguation) database. This database is supported by Klaus Tschira Foundation.
Step 2: Preprocessing
AND database used by proposed system is xml format. And for using it first it converted to structured format so that processing on it will be easy. For that xmltostruct() function of MATLAB is used. And this is termed as preprocessing of database.
Step 3: Feature Extraction
Third step is extracting feature which is given as input to clustering algorithm. So, author name is extracted from preprocessed database as it is provided as input to clustering algorithm. And after extracting, it is converted to its ASCII value as clustering algorithm works on numerical data.
Step 4: Classification
Fourth step classification i.e. clustering similar author names using multi-level clustering algorithm. Output of this step is displayed in graphical form.
Step 5: Output
In this step clusters generated in above steps are saved for testing.

Testing:

Proposed system is tested by testing input. First data to be tested is entered by loading data. After that data is preprocessed and author name is extracted from it. And disambiguity of it tested as is author name s ambiguous or disambiguous. And time required for checking ambiguity of input is also calculated.

III. RESULT ANALYSIS

Following table shows the comparison of input tested for AND database

Sr. No.	Test Input (Author Name)	Ambiguity (yes/no)	Time required for testing
1	Cao haong tru	No	0.053558
2	Haong Kiem	Yes	0.134564
3	Dinh Duy Li	Yes	0.012865
4	Dien Dinh	Yes	0.245615
5	Tu bao ho	No	0.12360

Table: Comparison of input tested for AND database

Above table shows the 5 different inputs testing and result produced. After testing input result is produced as is input ambiguous or disambiguous. For ambiguous result yes is written and for disambiguous result no is written. Also time required required for testing or executing input is also shown in table.

Total 30 inputs are tested for database out which 26 shows accurate result and hence accuracy of proposed system for AND database is 86.66 and error rate is 13.34.

IV. CONCLUSION

Here the system for author name disambiguation proposed to develop support to digital libraries and all its users. The efficient solution is developed with unsupervised classification techniques which are essential to get timely and satisfactory result. Here the system perform study of the various solution develop by different researchers. And one solution is proposed is a graph based multi-level clustering for disambiguation. The proposed multi-level graph based clustering approach takes the advantage of data mining techniques over traditional disambiguation techniques and tried to provide some effective solution. Using multi-level clustering the proposed system is trying to improve computational time and performance. This multi-level clustering approach for disambiguation has potential to develop excellent support for digital libraries.

V. ACKNOWLEDGEMENT

I would first like to thank Prof. Pravin Soni of P R Pote college of engg . My co guide Prof. Mayur S Burange was always willing to answer any query about my research on writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he taught I needed it. I would also like to thank the experts who were involved in validation survey for this research project. Without their passionate participation and input, the validation survey could not have been successfully conducted.

REFERENCES

[1] 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.
[2] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis.“Two supervised learning approaches for name disambiguation in author citations”,In JCDL '04: Proceedings of the 4th ACM/IEEE joint conference on

Digital libraries, ISSN No: 8179-6821p/p: 296—305, 2004.

[3] F. Wang, J. Li, J. Tang, J. Zhang and K. Wang. "Name Disambiguation using atomic clusters", In proceedings of the Ninth International Conference on Web –Age Information Management,p/p: 357-364,Washington,DC,USA,2008.

[4] X. Sun, J. Kaur, L. Possamai, and F. Menczer." Detecting ambiguous author names in crowdsourced scholarly data." In: IEEE third international conference on social computing (socialcom), ISSN NO: 8542-7422 p/p: 568—571, 2011.

[5] M. Wegmuller, J.P. Von Der Weid, P. Oberson and N. Gisin, "High resolution fibre distributed measurements with coherent OFDR", In: Proc. ECOC, p/p 109, 2009

[6] A. Veloso, A. A. Ferreira, M. A. Goncalves, H. F. A. Laender, and W. Meira Jr.,"Cost-effective on demand Associative Author Name Disambiguation", In: Information Processing and Management 48, p/p 680-697, 2012

[7] H. Peng, C. Lu, W. Hsu and J. Ho, "Disambiguating authors in citations on the web and authorship correlations", In: Expert Systems with Applications, p/p: 10521-10532, 2012

[8] H. Han, L. Giles, H. Zha, C. Li and K. Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations", In: Proc. Of Joint Conference on Digital Libraries, p/p: 296-305, 2004

[9] 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

[10] T. Masada, A. Takasu, and J. Adachi, "Citation data clustering for author name disambiguation", In: Proc. Of @nd International Conference on Scalable Information Systems, 2007

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

[12] Tasleem Sharif, "On the Use of Fuzzy Clustering in Name Disambiguation", In: Proc. of International Journal of Advanced Research in Computer Science", ISSN No: 0976-5697, p/p: 53-57, 2015

[13] Cota, R. G., Gonçalves, M. A., & Laender, A. H. F. . A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries. In Proceedings of the XXII Brazilian symposium on databases. João Pessoa, Paraiba, Brazil., p/p. 20–34, 2007

[14] Song, Y., Huang, J., Councill, I. G., Li, J., & Giles, C. L., Efficient topic-based unsupervised name disambiguation. In Proceedings of the 7th ACM/IEEE joint conference on digital libraries Vancouver, BC, Canada. p/p. 342–351, 2007

[15] Han, H., Giles, C. 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-CS Joint Conference on Digital Libraries. Tucson, AZ, USA, pp. 296–305, 2004

[16] Huang, J., Ertekin, S., and Giles, C. L. Efficient name disambiguation for large-scale databases. In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin, Germany, pp. 536–544, 2006

[17] Levin, F. H. and Heuser, C. A. Evaluating the use of social networks in author name disambiguation in digital libraries. Journal of Information and Data Management 1 (2): 183–197, 2010

[18] Tang, J., Zhang, J., Zhang, D., and Li, J. A unified framework for name disambiguation. In Proceeding of the 17th international conference on World Wide Web. Beijing, China, pp. 1205–1206, 2008.

[19] Culotta, A., Kanani, P., Hall, R., Wick, M., and McCallum, A. Author disambiguation using error-driven machine learning with a ranking loss function. In Proceedings of the International Workshop on Information Integration on the Web. Vancouver, Canada, 2007.

[20] Fan, X., Wang, J., Pu, X., Zhou, L., and Lv, B. On graph-based name disambiguation. ACM Journal of Data and Information Quality 2 (2): 10:1–10:23, 2011.

[21] Adriano Veloso, Anderson A Ferreira, Marcos Andr´e Gonc¸alves, Alberto HF Laender, and Wagner Meira. Cost-effective on-demand associative author name disambiguation. Information Processing & Management, 48(4):680–697, 2012.

[22] Mohammad Hossein Nadimi and Mostafa Mosakhani. A more accurate clustering method by using co-author social networks for author name disambiguation. Journal of Computing and Security, 1(4), 2015.