

Author Name Disambiguation using Web Information

KM. Pooja, Samrat Mondal, Joydeep Chandra
 Department of Computer Science and Engineering
 Indian Institute of Technology Patna
 Patna, India-801103
 Email: {pooja.pcs16, samrat, joydeep}@iitp.ac.in

I. BACKGROUND AND MOTIVATION

Author name disambiguation will help new researchers in expert finding. Author name disambiguation will do the correct attribution of records to the authors. Absence of unique identifier for an author puts an extra burden on the user to go through all the publications and check whether this is the publication of author he/she is looking for.

The famous on line tool DBLP even has not solve this problem till now and it merges the records of an author, for eg. in DBLP we have found more four or more different *Hao Wang's* publications' mixed in single Hao Wang profile as we can see in fig 1.

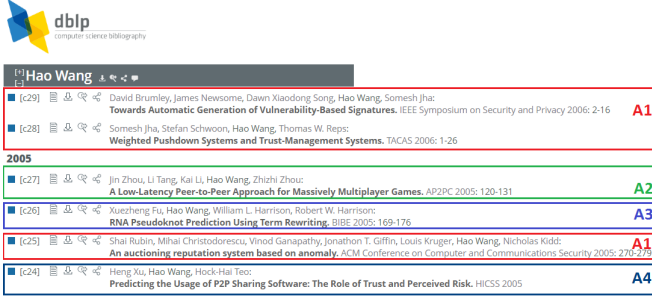


Fig. 1: Publications of 4 four or more Hao wang mixed in DBLP

II. COMPLEXITY OF PROBLEM

Traditionally, the problem of author name disambiguation has been addressed either using supervised learning or clustering based techniques. Apart from the author affiliations and coauthor set, external references like information from the web pages of the authors have also been considered as feature attributes. However, none of these features are found to be adequate in dealing with the problem. Further certain discriminative attributes are not available for all records, for example “Web Page” of every author is not available and if available, they are not complete. Some of the discriminative attribute like co-authors, affiliation, emails changes with the time.

III. PROPOSED APPROACH

There are two scenarios regarding publication pattern of an author. First : different authors with the same namesake may publish in similar domain. Second : an author may not only publish in same domain but also in different domains(areas). Existing solutions used to group records based on author's co-author network and topic itself. Existing solutions are mixing the records of different authors working in similar domain and fails to put together the publications of different domains. For capturing the first scenario we are using specificity of the problem along with the domain (when common co-authors are missing) itself, which reduces the chances of mixing the publications of different authors in same domain. For capturing the scenario when an author is working on different domain we are using web page of an author as auxiliary information.

For our experimental purpose we explore the following attributes of paper: “author name”, “co-author names”, “abstract”, “title of reference paper”, “author of reference” of dataset along with external information (web page of author). Work flow of our proposed method is shown in fig 2. “Topic Graph” is formed with abstract information and “Co-authorship Graph” is formed with the

help of co-authors information of papers. Common co-authors and similar topics provide a strong indication of a mapping between a paper and an author, so we form first level clusters $C_{f1}, C_{f2}...C_{fn}$ using these attributes. The similarity of domain and same author name in two different papers does not make it conclusive for them to map to the same author. So in this case we are exploring two more features, title of reference and author of reference of the given paper for specificity of the problem. With this information, we generate second level clusters $C_{s1}, C_{s2}, ..., C_{sn}$. Clusters formed with the specificity will decrease the chances of combining records of different authors with namesake. Publications of author's in a same or different domain can be found on author's web page. With using author's web page information along with second level clusters, we form final level clusters $C_1, C_2, ..., C_n$.

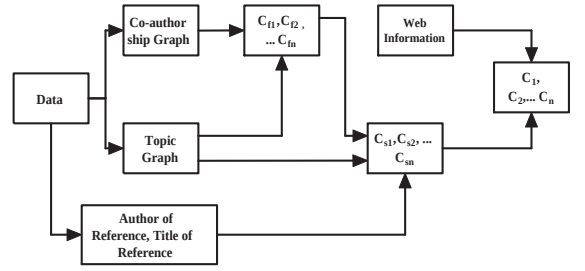


Fig. 2: Work Flow

PRELIMINARY RESULTS

We have experimented our method on a dataset of 13 authors and done comparative analysis with the state of art methods such as Hierarchical Heuristic Clustering (HHC) [1] and DBLP (online tool). After analysis we observe that our method has got significant improvement in terms of f_1 score as shown in fig 3. Improvement in f_1 score proves the efficiency of our heuristic. Incorporation of web information with heuristic enhances the paper grouping process.

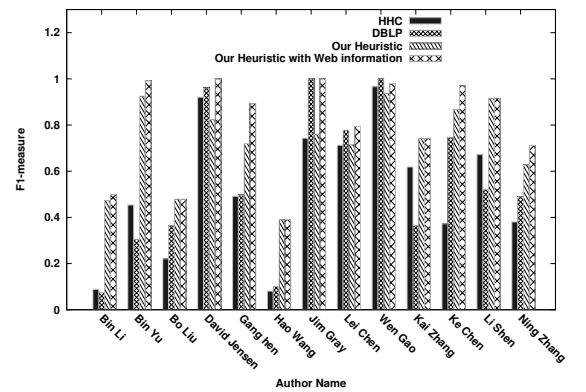


Fig. 3: Comparative results based on f1 score

[1] R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Goncalves, and A. H. Laender, An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations, Journal of the Association for Information Science and Technology, vol. 61, no. 9, pp. 18531870, 2010.