A Multistage Hierarchical Method for Author Name Disambiguation

Tasleem Arif\textsuperscript{a}, Rashid Ali\textsuperscript{b}, M Asger\textsuperscript{c}

\textsuperscript{a}Department of Information Technology, BGSB University Rajouri (Jammu and Kashmir)-185234 India, Contact: tasleem.ap@gmail.com

\textsuperscript{b}Department of Computer Engineering, AMU Aligarh, India.

\textsuperscript{c}School of Mathematical Sciences and Engineering, BGSB University Rajouri (Jammu and Kashmir), India.

Author name ambiguity has long been studied as a problem which affects literature management and leads to incorrect attribution of publications and credit to authors. Majority of the solutions provided either suffer from split citation problem or mixed citation problem. In recent years, there seems to be a tendency to use and store additional attributes of a publication to enrich its metadata. Use of e-mail ID of corresponding authors is prevalent in almost all publishing houses. In addition to the traditional metadata like author(s), title, venue and year, other attributes like e-mail ID and affiliation of author(s) are available in publication headers and in some cases in metadata also. For example, ACM Digital Library stores affiliations of authors as part of their metadata. In this paper, we propose a method that creates clusters in stages with each stage using a different attribute on the clusters created in the previous step. The purpose is to explore the effect of these additional attributes in resolving the author name ambiguity problem. Experimental results on publications obtained from DBLP show that our method obtained significant improvements over the existing state-of-the-art methods with average precision, recall and F-score of 93.93, 91.57 and 92.33 percent respectively and average execution time of 0.07 seconds per publication.

Keywords: Hierarchical Clustering, Metadata, Name Disambiguation.

1. INTRODUCTION

The advent of Internet has changed the way we live, communicate and maintain relationships. Research collaborations are not an exception and have benefited a lot from the advances in Information and Communication Technology (ICT) \cite{1} and web technologies have been deriving developments at the research front \cite{2}.

The rise in number of publications produced with each passing year has made it quite difficult for literature management services to properly index the publications of conflicting authors. This can be attributed to the fact that our parents had a limited number of options to choose our names from. In America alone 114 million males share 300 common names \cite{3} whereas in China the problem is more severe as 1.1 billion of its population shares just 129 surnames \cite{4}. If we look at the name ambiguity from the types of subproblem being faced in indexing \textit{i.e.,} split citation and mixed citation; split citation means that we have to manage publications of non-existent authors \cite{3} and mixed citations means that we are mixing two different real authors into one author. However the increase in percentage of joint publications can be treated as a relief for solutions to this problem as information about more authors is present in a publication metadata which leads to enrichment of publication metadata.

As pointed out by Liu \cite{5}, low availability of some key fields comprising of the metadata of a citation-record have resulted in poor performance of author disambiguation mechanisms. As such the need for extra information is being
very well. Experiments conducted and comparisons with other name disambiguation methods indicated substantial improvement over HAC and CONSTRAINT and considerable improvement over Fixed-K. Although the performance improvement over Fixed-K is only 1.35 percent, it is significant in the sense that it may become much more difficult to increase the performance more in the upper bracket of more than 90 percent prediction performance.

Publication venue title has been used for author name disambiguation purposes in previous studies as well as in this study but we are of the view that with the ever increasing publications it may not serve as a good feature for name disambiguation purposes as more and more similar authors may publish in same or relatively similar venues. The low recall and F1 score in case of Gang Wu can be primarily attributed with more than one Gang Wu publishing in similar publication venues.

An important observation that we made during this study was that in majority of the cases one author in a group of similar authors have almost more than forty percent of the publications. Another observation that we made during this study was that more the number of publications, the better the disambiguation results.

As a part of future work we intend to exclude the venue information and include title information and examine the effects of the intended change. In addition to that we also intend to make use of Soft-Computing techniques to deal with split-citation problem so that the precision may be increased further.

REFERENCES

7. Ferreira A A, Goncalves G A and Laender H F A. A Brief Survey of Automatic Methods

Table 4: Comparison of Precision, Recall, and F1 Scores (%age) of Representative Authors for Different Methods.

<table>
<thead>
<tr>
<th>Author</th>
<th>HAC</th>
<th>CONSTRAINT</th>
<th>Fixed-K</th>
<th>Our Approach-MSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Recall</td>
<td>F1</td>
<td>Prec</td>
</tr>
<tr>
<td>Jim Smith</td>
<td>92.43</td>
<td>86.80</td>
<td>85.53</td>
<td>70.91</td>
</tr>
<tr>
<td>Ajay Gupta</td>
<td>41.88</td>
<td>100.00</td>
<td>59.04</td>
<td>90.67</td>
</tr>
<tr>
<td>Michael Wagner</td>
<td>18.35</td>
<td>69.26</td>
<td>28.13</td>
<td>26.25</td>
</tr>
<tr>
<td>Rajesh Kumar</td>
<td>63.36</td>
<td>92.41</td>
<td>75.18</td>
<td>92.37</td>
</tr>
<tr>
<td>Hu Fang</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Jie Tang</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Richard Taylor</td>
<td>80.17</td>
<td>99.93</td>
<td>88.97</td>
<td>68.23</td>
</tr>
<tr>
<td>Paul Jones</td>
<td>36.56</td>
<td>89.00</td>
<td>50.00</td>
<td>38.64</td>
</tr>
<tr>
<td>Robert Fister</td>
<td>96.14</td>
<td>100.00</td>
<td>98.03</td>
<td>85.21</td>
</tr>
<tr>
<td>Gang Wu</td>
<td>97.54</td>
<td>97.54</td>
<td>97.54</td>
<td>71.86</td>
</tr>
<tr>
<td>Average</td>
<td>72.62</td>
<td>91.69</td>
<td>78.64</td>
<td>74.41</td>
</tr>
</tbody>
</table>
for Author Name Disambiguation, ACM SIGMOD, pages 15–26, 2012.


**Tasleem Arif** obtained his MCA from University of Jammu in 2004. He obtained his Ph.D in Computer Science from Baba Ghulam Shah Badshah University Rajouri in 2015. His Ph.D work was on Academic Social Network Extraction from online sources. He has authored about 25 papers in various International Journals and International conference proceedings. Currently he is working as Sr. Assistant Professor in the Post Graduate Department of Information Technology, Baba Ghulam Shah Badshah University Rajouri, Jammu and Kashmir. His research interests include Web Mining, Soft Computing, Information Retrieval, Data Mining and Cryptography and Network Security.

**Rashid Ali** obtained his BTech and MTech from AMU Aligarh, India in 1999 and 2001 respectively. He obtained his Ph.D in Computer Engineering in February 2010 from AMU Aligarh. His Ph.D work was on Performance Evaluation of Web Search Engines. He has authored about 75 papers in various International Journals and International conference proceedings. He has presented papers in many International conferences and has also chaired sessions in few International conferences. He has reviewed articles for some of the reputed International Journals and International conference proceedings. He has supervised 15 MTech dissertations. Currently, he is supervising four Ph.D candidates. His research interests include Web Searching, Web Mining, soft Computing (Rough-Set, Artificial Neural Networks, Fuzzy Logic etc.), and Image Retrieval Techniques.

**Mohammed Asger** obtained his M.Sc and Ph.D from Jamia Millia Islamia, New Delhi. He has vast teaching and administrative experience and has been instrumental in setting up of few Engineering colleges in the NCR region. He has authored about 50 papers in various International Journals and International conference proceedings. Currently he is working as Dean School of Mathematical Sciences and Engineering, Dean of Students and Principal University College of Engineering and Technology, Baba Ghulam Shah Badshah University Rajouri, Jammu and Kashmir. His research interests include Soft Computing, Quantum Computing, etc.,