Leveraging ResearchGate for the purpose of Author Name Disambiguation

In the light of the limitations of peer review approach, individual based bibliographic analysis plays an important role on the effectiveness of research evaluation (D'Angelo et al., 2011). In this context, identifying true authors for each publication is crucial to get precise and reliable results. This points to the problem of author name disambiguation (AND) which has been a fundamental issue in information science. Two major challenges of AND process are that, first, same author may write different papers under distinct names, second, distinct authors may have similar names. Besides, metadata present in databases is mostly short of necessary data. For example, as our main source, Thomson Reuters Web of Science database (WoS) indexes e-mail addresses and researcher ID (RID), a unique author identifier, which helps the process a lot (Ferreira et al., 2012). However, 13.79% and 8.57% are the percentages of the authors in our database having either e-mail or RID information, respectively.

Considering above-mentioned challenges, many approaches have been developed, most of which leverages machine learning models (Smalheiser and Torvik, 2009). While those machine learning methods have been applied by using author names and publication and journal (conference) names in most of the cases, some researchers have exploited external data through the Web and experienced a positive effect on AND process (Kanani et al., 2007; Yang et al. 2008). However, the main challenge when employing external data is the cost of extracting needed information from Web sources (Ferreira et al., 2012).

In this study, we leverage ResearchGate (RG), which is a social network site for scholars. In RG, subscribers can create their own profiles including publication lists. RG also provides publication lists for those non-members by collecting and grouping information from the Web. Being aware of possible drawbacks of RG, we suggest the use of RG along with some clustering results aiming at both confirming RG and automated results and checking how sufficient RG data is.

Methodology

RG, as the external data source is used in a two stage cleansing algorithm where we first use data indexed in the database to retrieve possible sets of papers linked to the same author. These assignments are then subject to a confirmation stage using RG. Firstly, we create a document network where the nodes refer to distinct authors of each of the separate papers. Links between the nodes indicate a possible match between two authors. These possible matches are deduced from similarity of author names, email address, affiliation, co-authors and Researcher ID. A weak connected component (CC) partitioning is applied to create groups of distinct authors which might refer to the same researcher. As this is a greedy algorithm the risk for false positives cannot be neglected. However, a CC is sufficient as edge cuts should not be issued in this stage of the procedure (false negatives). Other partitioning methodologies like modularity based community detection might deteriorate the assignment as they suffer from resolution limit which also introduces false positives without the reduction of false negatives.

Secondly, these groups of papers associated to individual researchers are matched against RG in order to confirm the assignment. Titles from publications are searched within RG. Top five results are retrieved and matched against the bibliographic data of the source publication using a character 3-gram text matching approach (Abdulhayoglu and Thijs, 2013). If a title match is found, the RG author profile URLs found under the returned title are scraped and extracted. From those RG author URLs, full names of authors are extracted and matched with the original source author name using a similar character 3-gram approach. In case of a confirmed match, the URL is assigned to the author as distinct identifier. The results from this second step are then used to approve or disapprove the partitioning of the first step.

Data

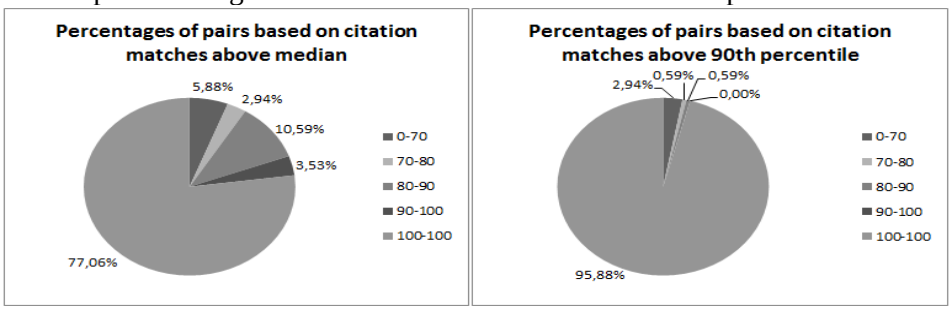
As illustration of this methodology is applied on a data set of 10,940 publications from our own field (bibliometrics) and their 31,983 authors found in WoS.

Results and Conclusion

The first step resulted in 16,236 (73.40%) distinct components. The size distribution is very skewed. Only 183 researchers have associated at least 10 papers. After manual inspection of these components we detected indeed several cases of false positives. Papers of author with very similar names are indeed grouped together. It is obvious that common names are more prone to this problem, especially Asian names.

Even though, in general, the overlap is quite high, there are some clusters where it is so low. When we check them in depth, we see that CC might act greedy and put irrelevant authors in a same group due to identical last names or identical e-mail records which are wrongly indexed in WoS. On the other hand, RG might lack of some records where CC grabs correctly. Under such facts we check the pairwise F scores. We get an F score of 0.955 when only CC is applied while it is 0.947 when mutual CC and RG results are used. Besides, we also check the citations received by the publications (especially highly cited ones) confirmed by mutual CC-RG results. We observe that the found and confirmed publications in RG are the ones having more citations. Figure 1 summarizes this finding.

Figure 1: The success of overlapping results based on joint results from CC clustering – RG when they are compared to original citation data above median and 90th percentile on individual basis.



For each author cluster, we assign them into 1of 5 formed groups based on their overlapped citation data with the original cluster. We implement this assignment only for those publications cited more than median or 90th percentile value of the original clusters. CC-RG results seem quite successful in that they reach all (100-100 group) the highly cited papers for 95.88% of all the clusters. As a result, we experience that RG contains a valuable data for the purpose of AND. By applying it, manual work required for AND process might be eased and it might provide more reliable data by confirming results retrieved by some unsupervised techniques.

References

Abdulhayoglu M.,A. & Thijs B. (2013). Matching bibliometric data from publication lists with large databases using N-grams. *14th International Society of Scientometrics and Informetrics Conference (ISSI-2013). Vienna, Austria, Proceedings*, 2, 1151-1158.

D'Angelo, C. A., Giuffrida, C., & 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), 257-269.

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*,41(2), 15-26.

Kanani, P. H., McCallum, A., & Pal, C. (2007). Improving Author Coreference by Resource-Bounded Information Gathering from the Web. *In IJCAI*, 429-434.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology*, 43(1), 1-43.

Yang, K. H., Peng, H. T., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2008). Author name disambiguation for citations using topic and web correlation. *In Research and advanced technology for digital libraries*, 185-196. Springer Berlin Heidelberg.