A Content and Social Network Approach of Bibliometrics Analysis across Domains

Christopher C. Yang
College of Information Science and Technology
Drexel University
chris.yang@drexel.edu

Xuning Tang
College of Information Science and Technology
Drexel University
xt24@drexel.edu

ABSTRACT

Bibliometrics data contain rich co-authorship network, text and temporal information. In this work, we employ a hybrid approach that incorporating content and social network similarity to conduct a bibliometrics analysis across the information retrieval and World Wide Web domains using the DBLP dataset.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]

General Terms

Measurement, Experimentation.

Keywords

Bibliometrics, Co-authorship Network, Content.

1. INTRODUCTION

Social computing has drawn substantial attention in the recent years. One important reason is the popularity of online social media in the new era of Web development, which open up a lot of opportunities for studying the online social behavior. In light of the recent development of social computing, some social computing techniques can be applied on bibliometric studies, which to some extent share common characteristics with the studies in social media (Tang and Yang, 2011a; Tang and Yang, 2011b; Yang and Tang, 2011). Citation networks and co-authorship networks are two typical networks that can be extracted from the bibliometric data and employed in bibliometric studies (Ding 2011; Garfield, 1972; Small, 1973). Citation network reflects how publications make citation to each other. A node corresponds to a publication and an edge represents a citation. However, citations occur without the authors knowing or working with each other (Lin et al., 2005). In addition, citations can span across a long period of time (i.e. an author may cite a paper written a few decades ago). As a result, the social relationship is much weaker in citation network. On the other hand, co-authorship network has been widely used to study the structure of collaborations and the status of individual researchers, since the collaboration relationships among users within a co-authorship network are more direct and convincing (Liu et al., 2005; Newman, 2004). However, only until recently, there are a few works which were devoted to study the dynamic feature of co-authorship networks. Barabasi et al. first employed empirical metrics to uncover the topological features of the coauthor networks at a given moment, and then tracked the time evolution of these quantities (Barabasi et. Al., 2002). They also inferred the structural mechanisms that govern these evolutions. Borner, Maru and Goldstone introduced a general process model that simultaneously governs the growth of co-authorship and citation networks (Borner, Maru, & Goldstone, 2004). In this work, we are interested in analyzing the dynamics of co-authorship networks which reflect how collaboration relationships change over time across different domains. In addition, bibliometric data offers a lot of content (text) information, which can be either the titles or abstracts of publications depending on their availabilities in a specific bibliometric dataset. This information supports the comparison of the topics covered across domains.

In this work, we employ a hybrid approach to study both content similarity and co-authorship network similarity simultaneously across different domains. It helps us to understand the similarity of collaborative groups in different domains. The high similarity between the co-authorship networks of two domains implies that there are collaborative groups participating in the scientific work on both domains. These two domains are likely to have work of common interests and/or highly relevant topics that are contributed by the similar groups of researchers. On the other hand, the content similarity indicates how similar two domains are based on the titles and/or abstracts of their publications. We extracted a dataset from DBLP data in the information retrieval and World Wide Web domains and conducted the hybrid approach of biobibliometrics analysis. The result showed that the two domains have an increasing trend of similarity.

2. PROBLEM DEFINITION

We first introduce the definition and notations used in this work and then define the research problem.

Definition 1 (Publication): A publication is an article published in a conference proceedings or a journal either in online or printed format. A publication has at least four attributes: title, year of publication, conference name (or journal name if it is published in a journal) and author list. Publication is the smallest unit in our study. We represent it by a tuple $t_i = \{W_{t_i}, N_{t_i}, opt_{t_i}\}$, where W_{t_i} is a TF-IDF term vector, $W_{t_i} = \{w_1^{t_i}, w_2^{t_i}, \ldots, w_{|t_i|}^{t_i}\}$, composed by terms from t_i's title and N_{t_i} is the coauthor network (defined below) associated with t_i and opt_{t_i} is the published year of t_i.

Definition 2 (Co-authorship Network): A coauthor network associated with a publication t_i is a fully connected graph $N_{t_i} = < V_{t_i}, E_{t_i} >$, where V_{t_i} is a set of authors, $\{p_1^{t_i}, p_2^{t_i}, \ldots, p_{|V_{t_i}|}^{t_i}\}$, coauthored in t_i, and E_{t_i} denotes the coauthor relationships between authors in V_{t_i}. Every pairs of authors in V_{t_i} are connected. As a result, the coauthor network of a publication t_i is a fully connected network.

iConference 2012, February 7–10, 2012, Toronto, Ontario, Canada.
ACM 978-1-4503-0782-6/12/02

3. HYBRID APPROACH INCORPORATING CONTENT AND SOCIAL NETWORK SIMILARITY

3.1 Content Similarity

As defined in Section 2, a publication is represented by a TF-IDF term vector and a domain is represented by the centroid of the collection of publications of this domain. Thus, the content similarity between two domains is the cosine similarity of their centroid term vectors, defined by $\cos(W_{A_x}, W_{A_y})$.

3.2 Network Similarity

Although content-based similarity can be effective to capture the commonality between two domains, its performance can be weakened by the vocabulary differences between the two domains that we may observe in the sparsity of common terms in different domains. Co-authorship network similarity considers the social network properties of the common authors in two domains rather than the vocabularies used by the authors. We measure the co-author network similarity by considering the intersection of important authors involving in two domains.

$$Overlap\left(N_{A_x}, N_{A_y}\right) = \frac{\sum_{P_k \in N_{A_x} \cap N_{A_y}} \min(ss(P_k^{A_x}), ss(P_k^{A_y}))}{\sum_{P_k \in N_{A_x} \cup N_{A_y}} \max(ss(P_k^{A_x}), ss(P_k^{A_y}))}$$

where $ss(P_k^{A_m})$ represents the significant score of the author P_k in the domain A_m, $N_{A_x} \cap N_{A_y}$ denotes the intersection of authors in the corresponding co-authorship networks, and similarly $N_{D_x} \cup N_{D_y}$ denotes the union of authors in the two co-authorship networks. The larger the intersection of authors involving in both domains, the higher the value of **Overlap(•,•)** is. Moreover, if an author involves in both domains similarly, the smaller the difference between the significant scores of this author in the two co-authorship networks and the higher the value of **Overlap(•,•)** is. In this work, we employ the degree centrality to measure the significance score of an author in a co-authorship network.

3.3 Hybrid Similarity

We then combine these two similarity measurement and get the hybrid similarity score for two domains denoted as:

$$\varepsilon \times \cos\left(W_{A_x}, W_{A_y}\right) + (1 - \varepsilon) \times Overlap\left(N_{A_x}, N_{A_y}\right)$$

where ε is a weight factor.

4. EXPERIMENT AND CONCLUSION

We extracted the dataset of two domains, information retrieval (IR) and World Wide Web (W3), from DBLP. The information retrieval domain consists of publications from SIGIR, ECIR and TERC conferences between 1995 and 2010. The World Wide Web domain consists of publications from WWW, WSDM and HYPERTEXT conferences from 1995 to 2010.

It's important to note that a co-authorship network of a domain may not necessary be a network of single component. In other words, it can consist of multiple components. If authors have a broader collaboration in this domain, there may be more number of authors in each component. For similar reason, each component may also have more number of papers. In Figure 1, we plot author/component, paper/component respectively for IR and W3. Based on this particular dataset and result, IR researchers tend to have more collaboration than W3 researchers, especially after 2000.

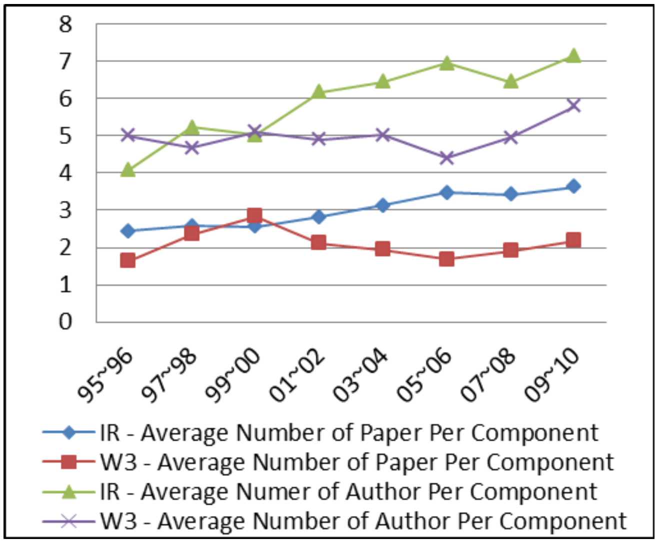

Figure 1. Co-authorship Network Statistics of IR and W3 Domains

We applied the hybrid approach on this dataset and the results are plotted in Figure 2.

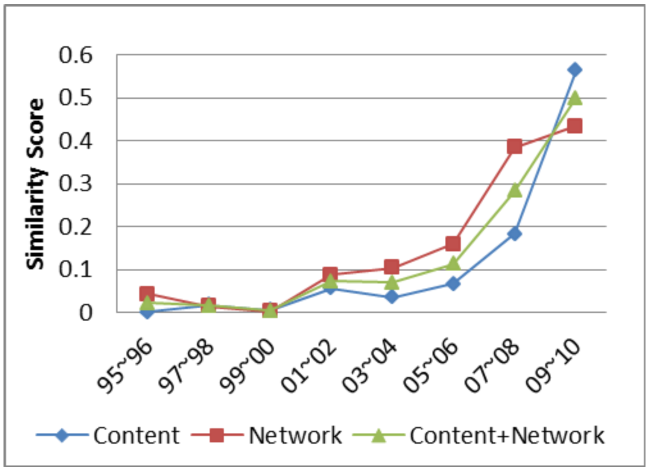

Figure 2. Content, Network and Hybrid Similarity Between IR and W3 Domains

In this figure, the green line demonstrates the similarity score calculated by our hybrid approach which measures the similarity between IR and W3 domains ($\varepsilon = 0.5$). As it is shown in figure 2, IR and W3 domains are becoming more similar in the last 15 years. In addition, we plotted the content similarity between these two domains in blue line ($\varepsilon = 1.0$) and the co-authorship network similarity in red line ($\varepsilon = 0$). Both of them show an increasing trend which also confirms the same observation. This observation is very reasonable because many information retrieval systems are now Web-based and many information retrieval techniques are developed for extracting Web pages or XML documents. At the same time, search engines and knowledge discovery are the most popular topics in W3 domain. This experiment shows the potential of our approach to study bibliometrics data by using content and co-authorship networks.

5. REFERENCES

[1] Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. 2002. Evolution of the social network of

scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3-4), 590-614

[2] Borner, K., Maru, J. T., & Goldstone, R. L. 2004. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5266-5273

[3] Ding, Y. 2011. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informatics*. *5*(1), 187-203.

[4] Garfield, E. 1972. Citation analysis as a tool in journal evaluation. Science. 178, 471-479

[5] Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. 2005. 2000. Co-authorship networks in the digital library research community, *Information Processing and Management*. 41, 1462~1480

[6] Newman, M. E. J. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5200-5205

[7] Small, H. 1973. Co-citation in scientific literature: New measure of relationship between two documents, *Journal of the American Society for Information Science*.

[8] Tang, X. and Yang, C. C. 2011a. Following the social media: aspect evolution of online discussion. *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, College Park, MD, March 29-31, 2011.

[9] Tang, X. and Yang, C. C. 2011b. Dynamic community detection with temporal Dirichlet process. *Proceedings of IEEE International Conference on Social Computing*, Boston, MA. October 9 – 11, 2011.

[10] Yang, C. C., Tang, X., and Gong X.. 2011. Identifying clusters from Dark Web with temporal coherence analysis. *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, Beijing, China, July 10-12., 2011.