SNDocRank: Document Ranking Based on Social Networks

Liang Gou[1], Hung-Hsuan Chen[2], Jung-Hyun Kim[2], Xiaolong "Luke" Zhang[1], C. Lee Giles[1,2]
Information Sciences and Technology[1], Computer Science and Engineering[2]
The Pennsylvania State University, University Park, PA, 16802, USA
{lug129, hhchen, jzk171}@psu.edu, {lzhang, giles}@ist.psu.edu

ABSTRACT
To improve the search results for socially-connect users, we propose a ranking framework, Social Network Document Rank (SNDocRank). This framework considers both document contents and the similarity between a searcher and document owners in a social network and uses a Multi-level Actor Similarity (MAS) algorithm to efficiently calculate user similarity in a social network. Our experiment results based on YouTube data show that compared with the tf-idf algorithm, the SNDocRank method returns more relevant documents of interest. Our findings suggest that in this framework, a searcher can improve search by joining larger social networks, having more friends, and connecting larger local communities in a social network.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *relevance feedbacks, retrieval models, selection process.*

General Terms
Algorithms, Experimentation, Human Factors.

Keywords
Ranking, Social Networks, Information Retrieval, Multilevel Actor Similarity.

1. INTRODUCTION
Ranking algorithms play a critical role in searching engines and research continues on the investigation of ranking algorithms. Some algorithms are user-neutral and measure the importance and relevance of documents mainly based on the contents and relationships of documents. Some personalized ranking methods have been proposed to improve search results by including various types of user information [5]. However, these algorithms largely focus on local activities of searchers, and fail to embrace the large social contexts of searchers.

When users are engaged in social networks through services (e.g., Facebook, Flickr, and YouTube) to communicate and share information with their friends, family, and colleagues, their social networks may provide richer and more reliable clues about the purposes and interests of their information search. To leverage these social contexts of searchers, we propose a new framework for personalized ranking, Social Network Document Rank (SNDocRank), which considers a searcher's social network in ranking the relevancy of documents. The premise of our methodology is that: 1) users tend to friend with those who share common interests, and 2) users are more interested in information from friends than from others. We also propose a Multilevel Actor Similarity (MAS) method to efficiently calculate user similarity in large social networks.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. SNDocRank FRAMEWORK
SNDocRank is a framework to rank the relevant documents based on the actor similarity of a searcher and other users in a social network. In this section, we first present the concept of SNDocRank. Then, we introduce our multi-level actor similarity (MAS) method to reduce the time complexity in calculating the actor similarity in our ranking method.

2.1 SNDocRank
In the SNDocRank framework, the ranking function is a combination of the basic term-document similarity function, such as tf-idf [6] score, and social network actor similarity. SNDocRank first identifies the user who issues the queries, and ranks the search result based on the similarity scores with others in the user's social network. Thus, SNDocRank score is given by

$$SNDoc(v, t_i, d_j) = f(sim(t_i, d_j), S_{vu}),$$

where v is the current user, t_i is the term, d_j is the document, u the owner of the document, S_{vu} is the similarity value between user v and u in a social network, and f is any possible function.

2.2 Multi-Level Actor Similarity (MAS)
Our MAS method aims to enhance the accuracy of actor similarity measurement by considering the global structure information of a social network, and also reduce the complexity of similarity computation by hierarchical clustering.

MAS is calculated with the structural features of a social network, i.e. how actors are connected with each other in a social network. This approach includes three steps. First, it clusters and aggregates a social network at multiple levels based on the network structure with a fast community detection algorithm proposed by Clauset, et al. [1]. Then it applies the LHN vertex similarity [4] to the clustered networks at each level by considering the weighted edges among actors. A simple way to apply the LHN vertex similarity to a weighted network is to inverse the weighted edge values of the network and then to apply

the LHN vertex similarity algorithm. Finally, global similarity values are calculated crossing all levels.

The process of MAS can be illustrated in Figure 1. Suppose we have a social network in Figure 1a, and this network can be clustered hierarchically, shown in Figure 1b and 1c. In Figure 1b, Nodes 1, 2, and 3 are grouped into one cluster—an abstract node C_{11}, Nodes 4 and 5 into C_{12}, and Nodes 6 and 7 into C_{13}. Thus, we have a backbone network consisting C_{11}, C_{12}, and C_{13}, and the edges among them.

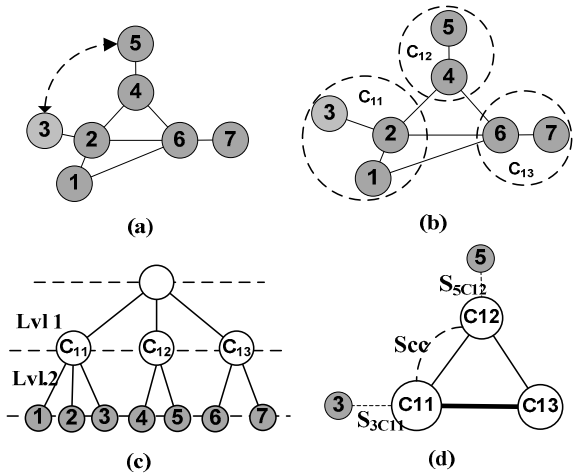

Figure 1. The Process of Multi-level Actor Similarity (MAS).

The hierarchical structure is shown in Figure 1c. To calculate the similarity between Nodes 3 and 5, which belong to C_{11} and C_{12} respectively, we first computed the similarity values between C_{11} and C_{12} (S_{CC}), between Nodes 3 and Cluster C_{11} (S_{3C11}), as well as between Node 5 and Cluster C_{12} (S_{5C12}) with weighted LHN, and then combine three similarity values together and get the final similarity value S_{3C11} S_{CC} S_{5C12}. Instead of computing the whole network in Figure 1a, we only consider the backbone network shown in Figure 1d. This approach offers a good scalability without scarifying the global structural information of the social network and can greatly reduce the computation complexity [2].

3. EXPERIMENTS

We implemented the SNDocRank framework in a mobile video social network application [3], and conducted experiments to evaluate this framework and the MAS method with YouTube datasets. The datasets include two fully connected social networks from YouTube—a larger social network that consists of 16,576 different registered users and their 39,281 videos; and a smaller network that has 2,264 users and 7,309 videos.

We first selected a set of users from two networks who we assumed need to search for videos in YouTube. Our assumption was that a returned result is good for a searcher only when the result is both relevant and interesting. The interest of a searcher is defined by the dominant category to which the videos that the searcher has uploaded belong. We chose three categories of interest – music, sports, and animation – and six different users with three levels of degree – high, medium, and low – in each category. We chose 15 queries related to each category.

To examine the effectiveness of the proposed ranking algorithm, we compared following three algorithms: a baseline condition with tf-idf [6]; the SNDocRank method with a cosine actor similarity algorithm in social networks; and the SNDocRank with our MAS method. We used a popular metrics, *Normalized Discounted Cumulative Gain (NDCG)* [6], to evaluate the ranking algorithms.

4. RESULTS AND DISCUSSION

The results of our evaluation studies indicate that overall, the SNDocRank framework can return better search results than the traditional tf-idf ranking algorithm in terms of document relevancy, the matching of interests with searchers, and the ranking effectiveness of returned results. Our MAS method outperforms the cosine similarity algorithm consistently across two social networks with different sizes, searchers with different degrees, and the interest groups with different sizes in a social network. This may imply that the structure of a searcher's social network can provide clues about the user's information needs and then can be used to improve ranking performances algorithms.

SNDocRank methods (both MAS and cosine) vary with the size of a searcher's social network, a searcher's degree, and the size of a searcher community in a social network. Although the SNDocRank method considers the global information of a social network, it becomes effective only as the size of a network reaches a certain magnitude. The degree of a searcher in a social network can affect the performance of the SNDocRank framework. Generally speaking, both MAS and cosine methods benefit high-degree searchers more than they do low-degree searchers. The size of local communities in a social network also affects the SNDocRank results. Both MAS and cosine algorithms favor larger interest groups.

Our results suggest some indicators that users can pursue to improve searching results. First, a user should join large social networks, because the SNDocRank method benefits large social networks more than small networks. Second, a user should try to be connected as many people as possible to increase the degree, which leads to better search results. Finally, in a social network, a user should be connected large communities or interest groups.

5. ACKNOWLEDGEMENTS

Part of this work has been funded by Alcatel-Lucent.

6. REFERENCES

[1] Clauset, A., M. E. J. Newman, & C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6), 66111, 2004.

[2] Gou, L., H. Chen, J. Kim, X. Zhang & C. L. Giles. SNDocRank: a Social Network-Based Video Search Ranking Framework. *Proc. of ACM MIR'*10, 2010

[3] Gou, L., J. Kim, H. Chen, J. Collins, M. Goodman, X. Zhang & C. L. Giles. MobiSNA: a Mobile Video Social Network Application. *Proc. of MobiDE'09*, 53-56, 2009

[4] Leicht, E. A., P. Holme & Newman, M. E. J. Vertex similarity in networks. *APS*, 73, 26120, 2006.

[5] Micarelli, A., F. Gasparetti, F. Sciarrone, & S. Gauch Personalized search on the World Wide Web. In *Lecture Notes in Computer Science*, 4321: 195-230, 2007.

[6] Salton, G. & C. Buckley. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5), 513-523, 1988.

[7] Jarvelin, K. & J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of SIGIR '00*, 41-48, 2000.