Topic-level Social Network Search

Jie Tang, Sen Wu, Bo Gao, and Yang Wan
Department of Computer Science and Technology, Tsinghua University
jietang@tsinghua.edu.cn, gb@keg.cs.tsinghua.edu.cn
{ronaldosen, wy_rdfz}@gmail.com

ABSTRACT

We study the problem of topic-level social network search, which aims to find who are the most influential users in a network on a specific topic and how the influential users connect with each other. We employ a topic model to find topical aspects of each user and a retrieval method to identify influential users by combining the language model and the topic model. An influence maximization algorithm is then presented to find the sub network that closely connects the influential users. Two demonstration systems have been developed and are online available. Empirical analysis based on the user's viewing time and the number of clicks validates the proposed methodologies.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining; J.4 [**Social and Behavioral Sciences**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

influence analysis, social networks, topic model, social search

1. INTRODUCTION

With the success of many large-scale online social networks (e.g., Facebook, MySpace, and Twitter), finding people and their connections is becoming one important issue. Given a network, an interesting question is: who are the most influential users on a specific topic? A further question is: how do the influential users connect with each other? Answering the questions can benefit many applications, for example, to understand how research ideas or innovations spread in an academic network. In this paper, the problem is referred to as topic-level social network search.

To quickly grasp the main idea, we give a motivating example in Figure 1 using a coauthor network. The

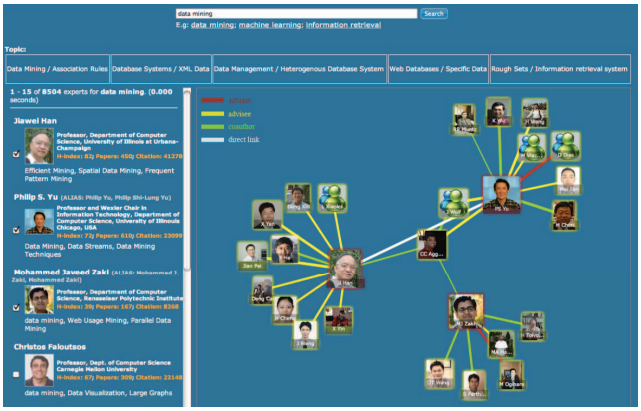


Figure 1: Topic-level social network search in coauthor network.

query is "data mining" and the system identifies five major relevant subtopics to the query, including "Data mining/association rule", "Database system", "Data management", "Web database", etc. The user can choose one of the subtopics to search for the most influential authors on the specific topic and to visualize the sub network that connects the influential authors. The colored links between authors indicate different types of relationships. For example, a red-colored link stands for an advisor relationship; while a yellow-colored link indicates an advisee relationship. With such a sub network, we can quickly get an insight of the core community on a specific (queried) topic.

The problem of topic-level social network search presents a set of challenges:

- First, how to identify the subtopics for a given query?
- Second, how to identify the most influential users on a subtopic?
- Third, how to find the sub network that closely connects the most influential users?

In this paper, employing the coauthor network as the basis in our experiments, we try to conduct a systematic investigation of the problem of topic-level social network search.Specifically, we employ an Author-Conference-Topic (ACT) model [10] to obtain the topic distribution of each author. On each topic, we combine the language model and the topic model to rank authors and select top ranked authors as the most influential authors. Finally, we present

KDD'11, August 21–24, 2011, San Diego, California, USA.

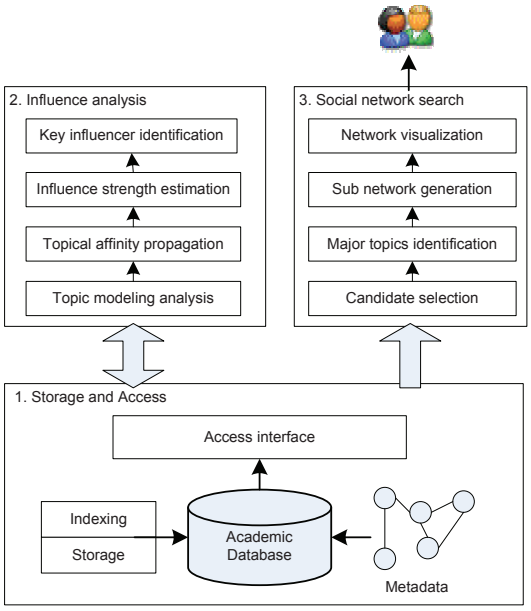


Figure 2: Architecture of topic-level social network search.

an influence maximization-based algorithm to find the sub network that connects the most influential authors for each topic. Based on the proposed approach, we have developed a prototype system[1]. The system has been online for nearly one year and attracted tens of thousands of users. Our empirical evaluation based on the user's viewing time and the number of clicks demonstrates the effectiveness of the proposed methods.

2. TECHNICAL SPECIFICATION

2.1 Architecture

Figure 2 shows the architecture of our approach to the topic-level social network search problem. In general, it consists of the following components:

- *Storage and Access*: it provides storage and indexing for the academic network data. Specifically, the academic database which consists of researcher profiles, publication papers, and conferences is constructed using a combination approach [8]; for storage it employs MySQL; for indexing, it employs the inverted file indexing method.
- *Influence analysis*: it performs an offline topic-based influence analysis. In particular, a topic modeling approach is used to find the topic distribution of all nodes and a topical affinity propagation is performed to quantify the influence between authors [7]; finally the influence strength and the key influencer (the most influential users) are identified based on the learned influence model.
- *Social network search*: Given a query, the system first ranks and retrieves authors as candidate influential authors using a method combining topic model and language model [10]; and then identifies which subtopics are relevant to the query and for each topic, it generates a sub network which centers with the most influential authors.

[1]http://arnetminer.org/association.do?m=home

2.2 Data Preparation

The academic data is located in the distributed Web. For searching and mining the academic network, we need first extract the networking data from the Web.

Some of the academic data can be extracted from structured data sources such as the publication information from DBLP; while other data needs to be extracted from unstructured Web pages such as researchers' homepages. We use a unified approach to extract researcher profiles from the researchers' homepages. We integrate the publication data from online databases. We extract the organization information from Wikipedia using regular expressions.

We employ a unified approach for researcher profiling [8] and a probabilistic approach for dealing with the name disambiguation problem in the integration [5]. The unified approach for research profiling explored in this paper is based on a new Condition Random Field model called Tree-structured Conditional Random Fields(TCRFs) [6].

The extracted/integrated data is stored into an academic network base. With the profiling and integration methods, we have collected 1,483,246 researcher profiles, 2,858,504 publications, 7,042 conferences, and 32,215,473 paper-paper citation relationships, 47,443,857 coauthor relationships, and 14,720,130 paper-published-at relationships. Based on the academic network, we have developed an academic search system, called Arnetminer.org. Services such as expertise search, course search, organization search, and topic browser have been provided. The system is in operation on the internet since 2006 and has attracted millions of users from 202 countries.

2.3 Topic-based Influence Analysis

To obtain the topic model for each node, one can use either the conventional topic model (e.g., pLSI or LDA), or a model combining the heterogeneous objects (e.g., ACT), or a model further combining the link information. In this paper, we employ the Author-Conference-Topic (ACT) model [10]. The model simulates the process of writing a scientific paper using a series of probabilistic steps. In essence, the topic model uses a latent topic layer as the bridge to connect the different types of objects. More accurately, for each object it estimates a mixture of topic distribution which represents the probability of the object being associated with every topic. For example, for each author, we have a set of probabilities $\{P(z_i|a)\}$, respectively denoting how likely author a is interested in topic z_i.

We use Gibbs sampling for parameter estimation. During parameter estimation, the algorithm keeps track of a $A \times T$ (author by topic) count matrix, a $T \times V$ (topic by word) count matrix, and a $T \times C$ (topic by conference) count matrix. Given these three count matrices, we can easily estimate the probability of a topic given an author $P(z|a)$, the probability of a word given a topic $P(w|z)$, and the probability of a conference given a topic $P(c|z)$. Interested reader can refer to [10] for more details.

Based on the network and topic distribution on the nodes, we formalize the social influence problem in a topical factor graph model and propose a topical affinity propagation on the factor graph to automatically identify the topic-specific

social influence [7]. The main idea is to leverage an affinity propagation at the topic-level for social influence identification. The approach is based on the theory of factor graph [4], in which the observation data are cohesive on both local attributes and relationships. In our setting, the node corresponds to the observation data in the factor graph and the social relationship corresponds to edge between the observation data in the graph.

After topical affinity propagation, we can obtain a pairwise influence score $\{\mu_{st}^z\}$ and the topic distribution $\{\theta_v z\}$, where $\{\mu_{st}^z\}$ indicate the influence of node v_s on node v_t for the topic z and $\{\theta_v z\}$ represents the probability of topic z given node v. Then we can generate the topic-level social influence graphs. Specifically, for each topic z, we first filter out irrelevant nodes, i.e., nodes that have a lower probability than a predefined threshold. An alternative way is to keep only a fixed number (e.g., 1,000) of nodes for each topic-based social influence graph. (This filtering process can be also taken as a preprocessing step of our approach, which is the way we conducted our experiments.) Then, for a pair of nodes (v_s, v_t) that has an edge in the original network G, we create two directed edges between the two nodes and respectively assign the social influence scores μ_{st}^z and μ_{ts}^z. Finally, we obtain a directed social influence graph G_z for the topic z.

2.4 Social Network Search

For a given query, we first retrieve candidate influential users. This can be done using a retrieval method. In this paper, we use a method by combining the language model and ACT model [9]. The relevance score of an author a to the query q is calculated by:

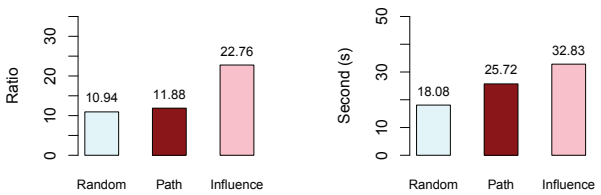
$$P(w|a) = P_{LM}(w|a) \times P_{ACT}(w|a) \quad (1)$$

where $P_{LM}(w|a)$ is the probability of generating word w from author a using the language model and $P_{ACT}(w|d)$ is the generating probability using the ACT model. Then those authors with a higher relevance score (e.g., larger than a threshold) are chosen as candidates.

As the second step, we identify major subtopics for these candidate authors. From the first step, we have a set of candidate authors A_q, and every author has a topic distribution $\{P(z|a)\}_z$. By accumulating the topic distribution of all authors in A_q, we will have a distribution of $\{P(z|A_q)\}z$. Then we rank $P(z|A_q)$ according to the probability and select five topics with the highest probabilities as the major subtopics.

Next, we generate the influence-based sub network. The algorithm is based on the notion of influence maximization [1, 3]. From the previous steps, we obtained a ranked list of candidate influential authors for each topic. The goal of this step is to find a sub network that closely connects the top k influential authors for each topic. The sub network can consist of a number of authors who connect the top k influential authors. This problem has been also studied in [2]. In this work, we formalize the problem as finding a set of authors in the network that maximizes the spread of influence under certain models. The problem is NP-hard, which can be proved by a reduction to the Dominating Set Problem. We utilize an approximate influence maximization-based algorithm [1]. The algorithm uses node's degree as the criterion to select nodes in the social network. To avoid redundant nodes (authors are close with each other in the network), we adopt a degree discount method. The idea is as follows: let v and u are neighborhood nodes in the network. If u has been selected in the graph, then when counting the degree of node v, we subtract the edge between u and v. To generalize it, when selecting any new node, we recount the degree of each node by subtracting the number of edges between the new node and any nodes already selected in the social graph.

Finally, we visualize the obtained sub network, which centers with the top k influential authors on the user selected topic.



(a) Expand/Remove ratio (b) Viewing time (s)

Figure 3: Empirical analysis.

2.5 Empirical Analysis

We have developed two systems to evaluate the presented algorithms. The first system is deployed in Arnetminer.org, an academic analysis and mining system [10] and the other system is deployed in Tsinghua university centenary celebration system[2].

Evaluation Criteria For evaluation, we consider the average user's viewing time and the average number of clicks. For user's viewing time, we mean when the user submits a query to the system, how much time he will stay on the returned social graph. Staying for a long time means that the user would be more interested in the result than that with a short time. We also design an interactive mechanism, which allows the user to expand an author's detailed social information when she is interested in knowing more about the person or to remove the node from the returned social network when she is not interested in the person. Finally, an Expand/Remove ratio is calculated. A high ratio indicates a better performance.

We compare our influence algorithm with two baseline methods: *Random*, which randomly selects authors to generate the sub network to connect the top k authors; and *Path*, which finds shortest paths between pairwise authors and combines the paths to generate the sub network.

Results Figure 3(a) shows the result of Expand/Remove ratio by different algorithms. It can be seen that our Influence algorithm has the largest ratio, while the Random and the Path algorithm have a lower ratio. Figure 3(b) shows the average viewing time of a user on the returned social graph by the different algorithms. It can be seen again that the Influence algorithm results in the longest viewing time, which confirms the effectiveness of our method. On average, the Influence algorithm almost doubles the ratio of Expand/Remove comparing with the two alternative algorithms; and gains an 27.64%-81.58% increase in terms of viewing time than the two baseline algorithms.

[2]http://alumni.cs.tsinghua.edu.cn/

3. DEMONSTRATION PLAN

We will present our system thoroughly in our demo. Particularly, we will focus on the following aspects.

1. First, we will use a poster to give an overview of the system, including the motivation and major issues addressed in the system. We will introduce the architecture and the main features of the system.

2. Next, we will explain how we conduct the topic-level influence analysis on the heterogeneous academic network. We will describe in detail the principled approaches for quantifying the topic-level influence between users.

3. Then, we will describe the principles for social network search. We will explain how we perform the major topic identification, sub network generation, and network visualization.

4. Finally, we will show statistics of system logs and user feedbacks. We will share our thoughts about the strength and the weakness of the system. We will further discuss the future work of the system.

Finally, please note that this is an ongoing project. Visitors should expect the system to change. We are extracting more researcher profiles and publications and are also developing more practical search services based on feedbacks from users.

3.1 What is New?

The prototype has been integrated into the Arnetminer.org system. In this subsection, we clarify what are new features of this demo, which are also the main contribution of this paper.

1. We present a novel influence analysis problem and introduce how to estimate the topic-level influence strength between nodes in a heterogeneous network. Based on the learned influence model, we identify the key influencer (the most influential nodes) on a specific node from the network. For instance, this demo will show you who are the most influential authors in "data mining", "database", or "machine learning" and who are mainly influenced by these key influencer.

2. We provide a novel social network search function based on the learned influence model. For a given query, the system automatically identifies which topics are the most relevant and which nodes are the relevant to the query and the topic. Then a topic-based sub network is generated according to the learned influence model.

3. We develop a web service interface for the influence-based analysis and search function, based on which the user can retrieve the topic-based sub network for a given query.

4. CONCLUSION AND FUTURE WORK

In this paper, we present a novel system of topic-based social network search. We study this problem on the heterogeneous academic network which consists of authors, conferences, and papers. We design an architecture to perform the influence analysis and topic-based social network search. We employ a topical affinity propagation approach to estimate the influence between nodes in the heterogeneous network and identify the key influencer based on the learned influence model. Further we provide a social network search service based on the output of the topic-level influence analysis. For a given query, the system automatically finds the most relevant topics for each topic, identifies the most influential users on each topic, and generates the sub network that closely connects the most influential users.

The general problem of social network search represents a new and interesting research direction. There are many potential future directions of this work. One interesting issue is to combine the learned influence model with users' behavior to build an more accurate user behavior model. Another issue is to integrate some background knowledge to help identify the most influential users. For example the advisor-advisee relationship can obviously help quantify the influence strength between coauthors. We also consider applying the influence model to some other domains, such as Twitter and flickr, to further validate its effectiveness. Finally, as the scale of the network grows very quickly, how to scale up the approach to handle very large (billions-of-nodes) network is also a challenging problem.

ACKNOWLEDGMENTS

The work is supported by the Natural Science Foundation of China (No. 61073073 , No. 60973102), Chinese National Key Foundation Research (No. 60933013, No.61035004).

5. REFERENCES

[1] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD'09*, pages 199–207, 2009.

[2] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *KDD'04*, pages 118–127, 2004.

[3] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146, 2003.

[4] F. R. Kschischang, B. J. Frey, and H. andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE TOIT*, 47:498–519, 2001.

[5] J. Tang, A. C. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *IEEE TKDE*, 99(PrePrints), 2011.

[6] J. Tang, M. Hong, J. Li, and B. Liang. Tree-structured conditional random fields for semantic annotation. In *ISWC'06*, pages 640–653, 2006.

[7] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD'09*, pages 807–816, 2009.

[8] J. Tang, L. Yao, D. Zhang, and J. Zhang. A combination approach to web user profiling. *ACM TKDD*, 5(1):1–44, 2010.

[9] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. Topic level expertise search over heterogeneous networks. *Machine Learning Journal*, 82(2):211–237, 2011.

[10] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.