EOS: Expertise Oriented Search Using Social Networks

Juanzi Li, Jie Tang, Jing Zhang
Tsinghua University,
Beijing China, 100084
{ljz, tangjie, zhangjing}
@keg.cs.tsinghua.edu.cn

Qiong Luo, Yunhao Liu
The Hong Kong University of Science and Technology, China
{luo, liu}@cse.ust.hk

Mingcai Hong
Tsinghua University,
Beijing China, 100084
hmc@keg.cs.tsinghua.edu.cn

ABSTRACT
In this paper, we present the design and implementation of our expertise oriented search (EOS) system. EOS is a researcher social network system. It has gathered information about a half-million computer science researchers from the Web and constructed a social network among the researchers through their co-authorship. The relationship in the social network information is used in both ranking experts on a given topic and searching for associations between researchers.

Categories and Subject Descriptors
H.3.4 [Information Systems]: Information storage and retrieval-*search process*, H.4.m [Information System Applications]: Miscellaneous

General Terms
Algorithm, Experimentation

Keywords
Association Search, Expert Search, Social Network

1. INTRODUCTION
Expertise Oriented Search (EOS) aims at providing comprehensive expertise analysis from distributed resources. It is useful in many applications, for example, finding experts on a given topic, detecting the confliction of interest between researchers, and assigning reviewers to proposals. Expertise oriented search addresses two main challenges: 1) retrieving largely unstructured raw data from different sources, extracting and integrating semantics from the data; and 2) effectively and efficiently analyzing and mining the large amount of data.

Driven by such applications and challenges, we have developed EOS, an expertise oriented search system. Currently EOS contains the information of 448,289 computer scientists, including their profiles and publications. This information is gathered from the Web and is integrated into a social network after cleaning and annotation. EOS provides four kinds of expertise oriented search services: (1) searching for a given person's information, (2) searching for publications, (3) searching for experts on a given topic, and (4) searching for associations between two researchers. The construction of the social network and the later two kinds of search services are the key points of the paper. The framework of social network construction and the algorithms proposed in this paper can be applied in social networks of other domains.

Copyright is held by the author/owner(s).
WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

The rest of the paper is organized as follows. We present the social network construction in Section 2. We describe our algorithms about expert search and association search in EOS in Sections 3 and 4. We discuss our system implementation and experiments in section 5. We conclude in section 6.

2. SOCIAL NETWORK CONSTRUCTION
Formally, a social network can be defined as a graph $G = (V, E)$, where $v \in V$ represents a person in the social network and $e^t_{ij} \in E$ represents a relationship with type t between persons v_i and v_j. We use $PLI(v)$ to represent the person x's local information such as person's name, title, and affiliation, and use $PRI(v)$ to represent the relationships of person v with the other persons. In EOS, co-authorship is used as one type of the relationships.

The major problem in the construction of a social network is to identify and integrate the person profile from distributed sources on the Web [1]. As the data sources, in EOS, we used Web pages related to a person, XML files of DBLP, and the publication ranking list from Citeseer (http://www.citeseer.com). The construction of the social network consists of three steps: 1) identifying related Web pages, 2) person information annotation, and 3) constructing the social network among researchers.

From the DBLP data, we obtained a list of researcher names. We use each of the researcher names as a keyword to search in Google and select the top 20 returned pages. Next a SVM (Support Vector Machines) based classification model is employed to identify the *really* relevant Web pages which include the information to the person. In Web page annotation, we use a rule-learning based annotation tool [2] to annotate from the identified Web pages the person information (including title, affiliation, address, phone number, fax phone number, and email address). We also use a classification based method to annotate the researcher's picture. We obtain other information about the researcher (such as coauthors' names and his/her detailed publication information) from DBLP. After that, we merge the gathered information of two parts and build a profile for the researcher. So far, we use heuristic rules to deal with the problem of name disambiguation.

After that, we use co-authorship as the relationship between researchers and build the social network automatically. Currently, there are 448,289 persons and 725,655 publications have been gathered in EOS. Totally, there are 2,413,208 relationships between persons with 5.38 relationships for each person on average.

3. EXPERT SEARCH

Most current work of expert finding focuses on how to rank experts by using a collection of documents or using information in the web pages or within enterprise. In EOS, we intend to improve the performance of expert finding from social networks by proposing a relevancy propagation-based expert finding algorithm. The algorithm consists of two stages.

In the first stage, we calculate *rel_l*(*x, q*), the relevancy of a person to a topic *q*, based on his/her local information:

$$rel\_l(x,q) = \lambda \times rel(d(PLI\_A(x)),q) + (1-\lambda) \times \sum_{i=1}^{m} w(p_i) rel(d(p_i),q)$$

where $d(PLI\_A(x))$, $d(p_1)$, $d(p_2)$, …, $d(p_m)$ are m+1 virtual documents constructed from $PLI\_A(x)$ and $Pub(x)=\{p_1, p_2, \ldots, p_m\}$. $Pub(x)$ is a set of papers authored by person x, and *PLI_A(x)* is person x's local information excluding his/her publications. *rel_l*(*x, q*) is composed of two parts. $rel(d(PLI\_A(x)), q)$ denotes the relevancy of person x's local information to the topic q except for his/her publication information and $rel(d(p_i), q)$ denotes the relevancy of paper p_i to topic q. $w(p_k)$ represents the impact factor of the publication p_k. We set λ=0.5 and define $rel(d, p) = p(q|d)$. Let $q=\{t_1, t_2, \ldots, t_n\}$, t_i is a word in topic q. We define $p(q \mid d) = \prod_{i=1}^{n} p(t_i \mid d)$ based on the probability independent assumption, where $p(t_i|d)$ is the probability of t in document d.

In the second stage, we propagate the topic relevancy of one researcher to his/her related researchers using a propagation-based algorithm [3]. The relevancy is updated by:

$$rel(x,q)^{i+1} = rel(x,q)^{i} + \sum_{e \in PRI(x)} co(e) \times rel(x',q)^{i}$$

where, $co(e)$ is the propagation weight of edge e. $rel(x, q)^0$ is the relevancy value calculated in the first stage. x' is a person which has relationships with person x.

4. ASSOCIATION SEARCH

In association search, the main difficulty is the time efficiency of the algorithm. A social network usually contains hundreds of thousands of nodes and millions of edges, and the response time to an association search should be within a few seconds. We propose a two stage association search method to search for connections between two researchers.

Let *asso*(*s*, *t*) be an association search task where *s* is the source person and *t* is the target person. In the first stage, we use a heap-Dijkstra algorithm to quickly find the shortest paths from all persons $v \in V-\{t\}$ in the graph to *t* with a running time of $O((m+n)log(n))$ where *m* is number of edges and *n* is the number of nodes in the social network. In the second stage, we implement a straightforward *s-t* association enumeration algorithm using the depth first graph search. The algorithm extends an *s-s'* association to *t* along the relationship *e*. We constrain the length of an association to be less than a pre-defined threshold *max_length*. We tentatively set it to be 7 in our experiments which can reduce the computational cost.

5. SYSTEM AND EXPERIMENTS

We have implemented the proposed algorithms and developed the system EOS, which is available at http://www.arnetminer.net. Now, EOS provides four kinds of expertise oriented search services including person search, expert search, associate search and publication search. We tested the performance of EOS. The average time of extracting person information from the Web and DBLP varies from few seconds to half dozen seconds. In expert search, the average search time is 2-5 seconds testing on expert finding tasks on 10 topics. In association search, the average search time in the network is less than 3 seconds.

For expert search, we assume that an expert to a topic is often active in the committees of the top conferences and organizations in the topic. We chose 5 topics and collected 5 lists of experts from the committees as the standard evaluation metrics (http://keg.cs.tsinghua.edu.cn/project/PSN/dataset.html). We evaluated the proposed expert finding algorithm in terms of the measures in [4]. Experimental results show that: (1) Our method significantly outperforms the baseline method using traditional IR methods; (2) In almost all the evaluations, our method outperforms the method which makes use of only person local information.

The test data sets used in association search are people pairs randomly selected from the social network. The number of people pairs in each test set is 1000. Experiments show that our proposed method achieves a high efficiency in all of the association search tasks. Our approach can find associations in less than 3 seconds on most of the test sets. Brute force enumeration based baseline method uses nearly more than four hundred times of our approach.

6. CONCLUSIONS

In this paper, we present an expertise oriented search system. We introduce the search services provided in the EOS system and conduct experiments to evaluate its performance. The system has been in operation for several months. We will open EOS to researchers so that they might extract information, edit and modify the information, and perform the complex expertise oriented search in social networks.

As the future work, we plan to investigate the extraction of more complex relationships and mining issues on social networks with the complex relationships.

7. ACKNOWLEDGMENTS

This work has been supported by Natural Science Foundation of China (90604025)

8. REFERENCES

[1] Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A. P., Arpinar, I. B., Joshi, A., and Finin, T. 2006. Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In *Proceedings of the 15th International Conference on World Wide Web* (WWW'2006). 407-416.

[2] Tang, J., Li, J., Liang, B., Huang, X., Wang, K. 2005. iASA: learning to annotate the Semantic Web. *Journal on Data Semantics*, Vol(IV), 2005. pp. 110-145.

[3] Arasu, A., Novak, J., Tomkins, A., Tomlin, J. 2002. PageRank computation and the structure of the web: Experiments and algorithms. In *Poster Proceedings of the 15th International Conference on World Wide Web* (WWW'2002).

[4] Craswell, N., Vries, A.P.d. and Soboroff, I. 2005. Overview of the Trec-2005 enterprise track. *TREC 2005 Conference Notebook*, pp. 199-205.