Discovering the Staring People From Social Networks

Dewei Chen
Department of Computer Science and Technology
Tsinghua University
Beijing, China
chendw@keg.cs.tsinghua.edu.cn

Jie Tang
Department of Computer Science and Technology
Tsinghua University
Beijing, China
jietang@tsinghua.edu.cn

Juanzi Li, Lizhu Zhou
Department of Computer Science and Technology
Tsinghua University
Beijing, China
ljz@keg.cs.tsinghua.edu.cn
ndcszlz@tsinghua.edu.cn

ABSTRACT

In this paper, we study a novel problem of *staring people discovery* from social networks, which is concerned with finding people who are not only authoritative but also sociable in the social network. We formalize this problem as an optimization programming problem. Taking the co-author network as a case study, we define three objective functions and propose two methods to combine these objective functions. A genetic algorithm based method is further presented to solve this problem. Experimental results show that the proposed solution can effectively find the staring people from social networks.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation

Keywords

Social network, Staring people discovery

1. INTRODUCTION

With the flourish of the Web 2.0 applications, people are getting more and more interactive and sociable. Finding people with high authority and sociality, referred to as *staring people discover*, is very important. Previously, this problem has been separately studied. For example, expert finding [1] tries to discover experts on a given query. While staring people discovery aims to find the persons with not only extensive knowledge but also strong social links.

In this paper, we formally define the problem of staring people discovery and propose a novel method to solve this problem. Given the information of a set of persons, our method extract a sub set of these persons which can represent the main information of both the persons and the relationships between them. We call the extracted persons as *staring people*. The problem of staring people discovery is relevant to, but different from, graph summarization [4], where the goal is to generate an abstractive representation of the graph data.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. STARING PEOPLE DISCOVERY

Generally, the social network can be modeled as a graph[2], in which the vertices indicate persons and the edges indicate the relationships between vertices. In this paper, we take the co-author graph as an example. In a co-author graph, each vertex indicates an author, and each undirected edge indicates two authors collaborated on some papers. The goal of discovery the "staring authors" is to identify the authors who are the reliable and active researchers in some domain. That means they should have published many papers and collaborated with many other researchers as well. Given a domain, we use the number of publications in this domain and the number of all publications as the profile of an author. And use the co-author times as the weight of the edge. Figure 1 shows an example of a graph in "Data Mining" domain and the staring authors discovered. The circle nodes indicate the authors and only the red(dark) nodes are staring authors. The first number bound with the author is the number of publication about "Data Mining" and the second is the number of all publications. Table 1 lists some notations used in this paper.

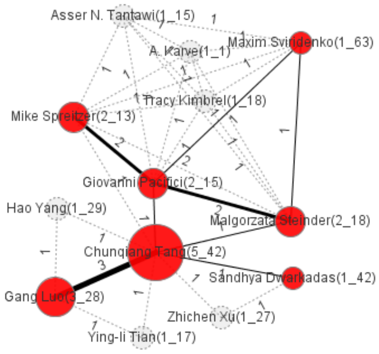

Figure 1: A co-author graph and discovery result.

Table 1: Notations

$G(V,E)$	a graph constructed by vertices V and edges E
$v_i \in V$	the i-th author
$e_{i,j \in E}$	the edge between author v_i and author v_j
n_i	the number of the author v_i's publications in domain
t_i	the number of all publications of author v_i
$c_{i,j}$	the co-author times of author v_i and author v_j
$D_{i,j}$	the nearest distance of author v_i and v_j (in spite of c)
α	the ratio of staring authors to all authors
$x_i \in \{0,1\}$	whether author v_i is a staring author.

By investigating the staring author discovery problem, we find that the following rules must be obeyed:

1. The number of selected staring authors should be as close to a predefined expectation as possible. The expectation can be either a fixed number or a ratio of string authors to all authors. In this paper, we use the second one and denote it as α.
2. The normal authors should be known by staring authors. Thus, The nearest distance between the normal authors and the staring authors should be as low as possible.
3. All authors must be treated equally during the discovery process.
   (a) If $n_i > n_j$, then author v_i should be selected as a staring author first.
   (b) If $n_i = n_j$ and $c_i > c_j$, then author v_i should be selected as a staring author first.
   (c) when $x_i = 1$ thus author v_i is a staring author, if $c_{i,j} > c_{i,k}$, then author v_j should be selected as a staring author first.

3. OUR APPROACH

The staring author discovery problem can be formalized as a programming problem which find a bitwise vector $(x_1, x_2, \cdots)$ best matching the previous rules. We can define a objective function for each rule, and then find the solution which optimizes the objective functions. The objective functions are:

$$\min f_1(x_1, x_2, \cdots) = \sqrt{\frac{10|\sum_i x_i - \alpha\|V\||}{\alpha\|V\|}}$$

$$\min f_2(x_i, x_2, \cdots) = \frac{\sum_{x_i=0} \min_{j\neq i\ and\ x_j=1}(D_{i,j}-1)}{\sum_i [x_i = 0]}$$

$$\min f_3(x_i, x_2, \cdots) = (1-\beta)\sum_{i\neq j\neq k}[x_i = x_k = 1 \neq x_j, c_{i,j} > c_{i,k}]+$$

$$\beta\sum_{i\neq j}[n_i > n_j, x_i = 0 \neq x_j] + [n_i = n_j, c_i > c_j, x_i = 0 \neq x_j]$$

where $[true] = 1$ and $[false] = 0$ and $\beta \in [0, 1]$.

Note that the objective functions are hard to be optimized simultaneously. We can either use multi-objective programming method or multilevel programming method to break this limitation. For multi-objective programming (MOP)[3], the objective functions are combined into a criterion function using the weighted sum method. And then, we aim to optimize the new criterion function. This solution can be formalized as:

$$\begin{cases} \min \sum_{i=1}^{3} \lambda_i f_i(x_1, x_2, \cdots) \\ s.t. \\ \quad x_i \in \{0,1\}, i = 1, 2, \cdots, \|V\| \end{cases}$$

The weight parameters λ_i can be set empirically or estimated by minimizing the error on a set of training data.

For multilevel programming (MLP)[3], we assume that the objective functions have different priorities. Thus, we can optimize the objective functions one by one according to the priorities. The MLP solution can be formalized as:

$$\begin{cases} \min f_1(x_1, x_2, \cdots) \\ s.t. \\ \begin{cases} \min f_3(x_1, x_2, \cdots) \\ s.t. \\ \begin{cases} \min f_2(x_1, x_2, \cdots) \\ s.t. \\ \quad x_i \in \{0,1\}, i = 1, 2, \cdots, \|V\| \end{cases} \end{cases} \end{cases}$$

Both MOP and MLP solutions are difficult to solve precisely. We propose to use the genetic algorithm to calculate the approximate results. In the genetic algorithm, the bitwise encoding method is used to encode a solution $x = (x_1, x_2, \cdots)$.

4. EXPERIMENTAL RESULTS

To evaluate our proposed methods, we generated 308 graphs according to the papers published from 2003 to 2008 in SIGKDD, SIGMOD, VLDB, etc. And three Ph.D. students annotated the staring authors in these graphs. The annotation results were combined by the voting method. Then, the combined result were used as the benchmark. Some of the them are publicly available. [1] And we use precision and recall to evaluate the discovery results.

We implemented the proposed method and conducted experiments on the 308 graphs. 50 of them were randomly selected as the training data. And the parameters β and λ_i were tuned on the training graphs. Then, with these parameters, we evaluated both MOP and MLP solutions on the other 258 graphs. The evaluation results are listed in table 2.

Table 2: Experimental results

Method	Precision			Recall		
	Max	Min	Average	Max	Min	Average
MOP	.9180	.6312	.8391	.8731	.7615	.8032
MLP	.9576	.7471	.8931	.9137	.8407	.8661

5. CONCLUSION

We investigate the problem of finding the most authoritative and sociable persons in social networks in this paper. More specifically, we use the co-author network as an example, and formalize this problem as two types of goal optimization programming problems. And we propose using the genetic algorithm to solve them. Experimental results indicate that our solution can properly identify the most representative and well-known authors out of the co-author network.

Please mention that the proposed solution is quite general. It can be applied to many other applications besides staring author discovery, such as staring blogger discovery, excellent survey paper discovery.

6. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of SIGIR'06*, pages 43–50, 2006.

[2] Y. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *Proceedings of WWW'08*, pages 685–694, September 2008.

[3] B. Liu. *Theory and Practice of Uncertain Programming.* Springer-Verlag, Berlin, 2009.

[4] S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In *Proceedings of SIGMOD'08*, pages 419–432, 2008.

[1] http://www.arnetminer.org/graph/review2.jsp