Social Network Structure behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track

Haiqiang Chen, Huawei Shen, Jin Xiong, Songbo Tan, Xueqi Cheng Institute of Computing Technology Chinese Academic of Sciences Beijing, China 100080 Email: chenhq@software.ict.ac.cn

Abstract— Expert finding system is a challenging problem in the enterprise environment. This paper introduce our research and experiments on TREC 2006's expert searching track. In our experiments, we find some interesting features of the community structures in the mailing list network. We also use some link analysis approaches to rank the candidates in the social networks. In our experiments, we choose the PageRank algorithm and a revised HITS algorithm as link analysis methods. These approaches give reasonable results in our experiments.

I. INTRODUCTION

Expert finding task is one of the two tasks in TREC 2006 Enterprise Track. The goal is to find a list of candidates who are experts in the specified topic. Our team participated in this task this year. This document is a brief report of our research and experimental results on expert search.

This is the second year of this expert finding task. In the first year of this task, most participants use some basic information retrieval models to solve this problem[1][2][3][4]. There are mainly two kinds of approaches: the two-stage model approach of combining relevance and co-occurrence, such as MSRA054[1], and the personal description file based approach, such as THUENT0505[3].

In this year, our team's research and experiments mainly focus on the mailing list corpus and the link relationship amongst the candidates expert and other users. The W3C corpus includes a large archive of the W3C's mailing lists. These lists are email forums for people who want to share information about W3C's research and projects. We can treat these forums as social networks.

In our experiments, we find some interesting features of the community structures of these networks: In most of the mailing lists, the candidate experts are not well connected. The social network in these mailing lists can be divided into some communities which includes a few candidate experts and a lot of other users. The candidate experts are mostly in the center of their communities.

And also, we use some link analysis approaches to rank the candidates in the social networks. In our experiments, we choose the PageRank[5] algorithm and a revised HITS[6] algorithm as link analysis methods. These approaches gives satisfying results in our experiments.

TABLE I W3C collection by scope

| Scope | Corpus size (gigs) | Docs | Avdocsize (kb) |
|--------|--------------------|---------|----------------|
| lists | 1.855 | 198,394 | 9.8 |
| www | 1.043 | 45,975 | 23.8 |
| esw | 0.181 | 19,605 | 9.7 |
| other | 0.047 | 3,538 | 14.1 |
| people | 0.003 | 1,016 | 3.6 |
| all | 5.7 | 331,037 | 18.1 |



Fig. 1. Degree distribution of the full mailing list network.

II. COMMUNITY STRUCTURE IN THE MAILING LIST NETWORKS

The W3C corpus includes a mailing list archive. As shown in Table I, the mailing list content (lists) is the largest portion of the entire corpus. An mailing list, is a special usage of email that allows for distribution of information to many email users. As for W3C, the mailing list is a type of Internet forum, in which users and experts can discuss anything within the open group.

The lists scope of the corpus includes mails, archive indexes, and other miscellaneous pages. In our research and experiments, we only focused on the mails. After parsing the lists scope, we found a total of 174311 mails in 251 lists. The users of the mailing lists forms a social network. In this network, each vertex is a user of the mailing list. If user A posts a mail to the mailing list, and user B replies that mail in the list, we put an edge from B to A. If user B posts a mail to the mailing list with receiver A, we also put an edge from B to A.

So we have a social network of the mailing list users. First we focused on the full network of all mailing list users using the W3C mailing list systems. It is shown in Figure 1 that this network is scale-free[7]. We try to calculate the clustering coefficient[8] of this network. The clustering coefficient of this network is $C \doteq 0.267$. It is much larger than the clustering coefficient of the comparative random network where $C_{random} \doteq 0.00041$. It is indicated that the mail network is highly clustered.

Assortative mixing is also an important feature in the complex networks. It is the extend to which nodes connect preferentially to other nodes with similar characteristics. In our research, the nodes of the network can be divided into two types: the candidate experts and the none-candidate users. So the assortative coefficient of the mail network is $r \doteq 0.148$. And we can find that candidate experts replies more often than non-candidate users. On average, one candidate expert replies 29.2 mails in the mailing list systems; and one non-candidate user replies only 6.9 (non-candidate users who never reply one mail in the mailing list system are not counted in).

The small assortative coefficient shows that this network is not very strongly assortative by the candidate/non-candidate partition. It implies that in this mailing list network, a candidate expert communicates more often with non-candidate users than with other candidate experts.

This phenomena (of relatively small assortative coefficient) can be explained as the usage of the mailing list. As we known, the 1092 candidate experts are all active participants in the W3C. Generally, they are more familiar with each other than with other non-candidate users. As for them, the W3C mailing list system is an open environment to communicate and discuss problems with other people outside the W3C organization. If one candidate expert want to discuss with another candidate expert, he (or she) may prefer personal mail to public mailing list. As for non-candidate users, usually they use the mailing list system to get some help from the experts. So they often have more discussion with candidate experts than with noncandidate users. As a result, the mail network is relatively disassortative by the separation of candidate experts and nonecandidate users.

This phenomena also implies that when the network is divided into communities, the candidate experts will be sparsely placed in many communities. Many networks, it is found, are inhomogeneous, consisting not of an undifferentiated mass of vertices, but of distinct groups. Within these groups there are many edges between vertices, but between groups there are fewer edges, producing a structure which is called the community. In our research, we use a fast community detection algorithm by Aaron Clauset[12] to detection the densely linked community in the mailing list network.



Fig. 2. The mail network in the www-annotation@w3.org mail list (only the largest connected component). Square nodes denote for candidate experts. Different colors imply the community structure in this network.

In order to take a closer view of the mailing list network, we choose the network in one relatively small mailing list. Figure 2 show the structure of the network in the www-annotation@w3.org mailing list. The community detection algorithm divides the network into 6 communities, each denoted by different colors. The 17 candidate experts (square nodes) in the network are sparsely placed in 5 communities. The largest community (of color blue) contains 8 candidate experts.

Most candidate experts in this network are in the center of the network. Table II show the top 10 closeness centrality[11] score of the network in www-annotation@w3.org. There are 6 candidate experts in Table II. In fact, we can find it easily that the other 4 non-candidate users in Table II are all experts in the given field (annotation service) even they are not in the candidate list.

Experiments on the networks of other mailing lists also show similar results.

So we can conclude from our experiments that: 1. The network in the mailing list is disassortative by the separation of candidate experts and non-candidate users; 2. Candidate experts do not gather into some densely linked groups (or cluster) in the mailing list network. They can be separated to some communities with other non-candidate users. 3. Candidate experts are more likely to be in the center of the mailing list network. They often have a larger closeness centrality score.

These conclusions will be helpful to the problem of expert finding.

TABLE II

TOP 10 CLOSENESS CENTRALITY SCORE IN THE NETWORK OF THE MAILING LIST WWW-ANNOTATION@W3.ORG.

| User id or mail address | Closeness centrality | |
|------------------------------|----------------------|--|
| marja@w3.org | 0.439252 | |
| candidate-0127 | 0.428571 | |
| candidate-0001 | 0.401709 | |
| matthew@mjwilson.demon.co.uk | 0.387363 | |
| candidate-0018 | 0.385246 | |
| candidate-0671 | 0.384196 | |
| candidate-1043 | 0.383152 | |
| candidate-1046 | 0.380054 | |
| julieg@weborganic.com | 0.379032 | |
| ping@lfw.org | 0.377005 | |

III. SOCIAL NETWORK LINK ANALYSIS IN THE MAILING LIST NETWORK

As we can see from the TREC 2005 proceedings, most group use an information retrieval model combining the relevance of documents and the co-occurrence of experts to solve the expert finding problem.

Our approach this year is totally different from this information retrieval model. First, we do a query in our document search engine on the W3C corpus(only mails content scope). It returns a list of relevant mails. We can extract the send/receive and reply relationships between all senders and receivers of the relevant mails. So we get a social network of the users discussing the given topic. Using link analysis and social network methods, we can get a ranked list of the users and candidate experts on the given topic.

As a result from our analysis in Section II, we can conclude that most candidate experts are in the center of the networks. So we first try to use the closeness centrality score to rank the candidate experts. But this method does not give a satisfying result. We also try the betweenness centrality[10] score to rank users and it fails again. These two methods can find out the top ranked candidate experts nicely, but often miss the other candidates.

Then we try to use other link analysis methods instead. The most well-known link analysis application is the web link analysis. And the most popular methods for web link analysis are PageRank[5] by Page and Brin, and HITS[6] by Kleinberg. We try these two methods in our experiments.

The PageRank and HITS algorithms are all for web link analysis. They all assume that the link between web pages implies some functional meaning. As for our application, we also assumed that: if user A post a mail to user B or reply a mail posted by B while the mail is relevant to topic Q, user B will recommend user A as an expert in the topic Q; so we can add a link from B to A. This assumption is quite reasonable for W3C mailing list network considering the result in Section II: the candidate experts reply more mails than non-candidate users; the candidate experts send / reply mails more often to

HITS vs. two-stage model



Fig. 3. MAP comparison between the two stage model and the HITS algorithm.

non-candidate users than to candidate experts.

We apply the PageRank algorithm to rank the users in the mail network. We also use a revised HITS algorithm to get a run result in our experiments. As a comparative group, we also set up a two-stage model combining relevance and co-occurrence. The HITS algorithm gives an relatively smaller average precision to the two-stage model as we can see from Figure 3.

But the two algorithms both failed to provide a satisfying result. An main reason is supposed to be the difference between the structure of the web and the mail network. In the assumption of the PageRank algorithm, the link from page A to page B is actively chose by A; but in the mail networks, the link from user A to user B is added when B answers A. It is different from the recommendation relationship implied by the web links. In Kleinberg's assumption in HITS algorithm, the web is supposed to have a hub-authority structure. But as for the mail network, one can hardly find such a hub-authority structure. If user want some help, they can directly connect to the experts. The proxy nodes(as the hub in the web) can not be found in mail networks.

IV. SUBMITTED RUNS

In this years, we submitted 5 runs for the expert search track. The 5 runs are as follows:

- *ICTCSXRUN01*: This run use a simple two-stage model combining relevance and co-occurrence. Only lists and www scope of the corpus is used. This run is used as a comparative run. This run get the best MAP and bpref score in our 5 runs.
- *ICTCSXRUN04*: This run is a tune of the ICTC-SXRUN01. A clustering-based re-ranking is used.
- ICTCSXRUN03: Based on the relevant document list provided by ICTCSXRUN01(mails only), we can construct a mail network of the users and candidate experts. This

TABLE III

SUBMITTED RUN RESULTS.

| RUN ID | MAP | bpref | P@10 |
|-------------|--------|--------|--------|
| ICTCSXRUN01 | 0.4949 | 0.4858 | 0.5837 |
| ICTCSXRUN02 | 0.4803 | 0.4715 | 0.5878 |
| ICTCSXRUN03 | 0.2005 | 0.2629 | 0.3612 |
| ICTCSXRUN04 | 0.4734 | 0.4634 | 0.5796 |
| ICTCSXRUN05 | 0.4219 | 0.4300 | 0.5347 |
| best | 0.7507 | 0.7542 | - |
| median | 0.3412 | 0.3602 | - |
| worst | 0.0024 | 0.0070 | - |

run uses the PageRank algorithm to rank the candidate experts.

- *ICTCSXRUN05*: Based on the top result of the ICTC-SXRUN04 as the root set, we use the HITS algorithm to re-rank the candidate experts.
- *ICTCSXRUN02*: This is a trial run. We use the result from ICTCSXRUN01 and ICTCSXRUN03, and make a linear combination to get a new result.

Table III presents the performance achieved by our submitted runs, along with that of the participants.

V. CONCLUSION

In our submitted runs, the two link analysis algorithms did not give satisfying results. But our experimental analysis about the community structure of the social network in the mailing list can give us a lot of helpful information to solve the expert finding problem in our future research.

REFERENCES

- Y. Cao, H. Li, J. Liu and S. Bao, *Research on Expert Search at Enterprise Track of TREC 2005*. NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), 2005.
- [2] J. He, C. Chen, C. Yao, P. Yin, and Y. Bao, *Peking University at the TREC 2005 Question and Answering Track*. NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), 2005.
- [3] Y. Fu, W. Yu, Y. Li, Y. Liu, and M. Zhang, *THUIR at TREC 2005: Enterprise Track*. NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), 2005.
- [4] N. Craswell, A.P. de Vries, and I. Soboroff, *Overview of the TREC 2005 Enterprise Track*. NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), 2005.
- [5] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. Technical report, Stanford Digital Library Technologies Project, 1998.
- [6] J.M. Kleinberg, Authoritative sources in a hyperlinked environment. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [7] A.L. Barabasi, R. Albert*Emergence of scaling in random networks*. Science, 286:509-512, October 15, 1999.
- [8] M.E.J. Newman, *The structure and function of complex networks*. SIAM Review, 45(2):167–256, 2003.
- [9] M.E.J. Newman, *Mixing pattern in Networks*. Physical Review Letters 89, 208701, 2002.
- [10] L.C. Freeman, A set of measures of centrality based on betweenness. Sociometry. 40:35-41, 1977.
- [11] G. Sabidussi, *The centrality index of a graph*. Phychometrika, 21:581-603, 1966.
- [12] A. Clauset, Finding local community structure in networks. Physical Review E 72, 026132, 2005.