2007 IEEE/WIC/ACM International Conference on Web Intelligence

Finding Experts Using Social Network Analysis[1]

Yupeng Fu, Rongjing Xiang, Yiqun Liu, Min Zhang, Shaoping Ma
State Key Lab of Intelligent technology & systems
Tsinghua University
Beijing, China P.R.
Yupeng.Fu@gmail.com

Abstract

Searching an organization's document repositories for experts is a frequently occurred problem in intranet information management. A common method for finding experts in an organization is to use social networks – people are not isolated but connected by various kinds of associations. In organizations, people explicitly send email to one another thus social networks are likely to be contained in the patterns of communication. Moreover, in some web pages, the relationship among people is also recorded. In our approach we propose several strategies in discovering the associations among people from emails and web pages. Based on the social networks, we proposed an expertise propagation algorithm: from a ranked list of candidates according to their probability of being expert for a certain topic, we select a small set of the top ones as seed, and then use the social networks among the candidates to discover other potential experts. The experiments on TREC enterprise track show significant performance improvement with the algorithm.

1. Introduction

When locating some desired information or source of knowledge, one can usually be satisfied by finding an expert in the topic of interest. Thus finding experts in a particular area is a frequently encountered problem within any commercial, educational or government organization.

Yet the very problem of discovering who knows what is often challenging, we could build up a relationship between a query and an expert via documents. Plus, social networks — relationships among people in an organization — provide another opportunity for finding experts. People is not isolated but connected. When we locate someone as an expert, those candidates who have strong relationship with him are much possible to be experts too.

In enterprise corporation there are a lot of sources of information can be utilized to mine the relationship among the people. Email is a valuable source of expertise according to the fact that email is used widely today as a major means of communication. Both the content of email and the patterns of communication contain information about who knows what in an organization. Web pages in the organizations provide another repository of information that can be utilized to build the social network. As such it contains precious information about events, activities, works and interests of individuals.

In our previous work [1] we proposed a method associating candidates with documents to quantitatively estimating the individual expertise. In this paper, we concentrate on estimating the associations among candidates to build a social network and introduce an expertise propagation algorithm to help re-rank the expertise probabilities of the candidates to the given query. Under the algorithm, a candidate can acquire extra expertise probability from a reliable expert who has strong relationship with the candidate. The goal of this paper described here is to examine the various kinds of associations among candidates.

2. Expertise propagation

The expert finding problem can be defined as what is the probability of a candidate c being an expert given the query topic q. However, directly estimating the probability $P(c \mid q)$ is not possible since what we have are merely the documents (Web-pages, Emails, etc). In [1] we describe the approach of estimating $P(c \mid q)$ according to the association between the candidates and the content of the documents, which we refer as stage1. There is another kind of association which is among the candidates themselves. People is not isolated but connected. The connection among candidates can be identified through document analysis, such as those members in a same interest group may communicate frequently through emails and the candidates who publish a technical report may appear as co-authors. The associations among candidates can help much on identifying experts according to the fact that given an expert c_x, the candidate c_y who has strong association $a(c_x, c_y)$ with him is also quite likely to be an expert.

Thus we introduce an expertise propagation process for finding experts. We employ the associations among

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108)，Natural Science Foundation (60621062, 60503064) and National 863 High Technology Project (2006AA01Z141)

0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.14

candidates to propagate the likelihood from those highly possible experts to other candidates. This process can be viewed as estimating the probability $P(c_y \mid c_x)$: what is the probability of candidate c_y to be expert given the expert c_x . We get a set S of N candidates who are the most likely to be the experts from the stage 1 as the seed. Then, expecting that those candidates who have strong associations with the experts are also experts, we assign the expertise score from the experts in S to those candidates.

As a matter of fact, for a candidate c_x in set S and a candidate c_y to be examined, the less possible that c_x is an expert and the weaker the association $a(c_x, c_y)$, the less certain we are that c_y is an expert. This fact suggests that we split the expertise probability as he propagates to other candidates according to their associations: if an expert c_x has an expertise probability of $P(c_x)$ and he has ω associated candidates, each of the ω candidates c_y has the association $a(c_x, c_y)$ will receive a score fraction from c_x . The fraction is

$$P(c_y \mid c_x) = \frac{a(c_y, c_x)}{\sum_{c_i \in \omega} a(c_i, c_x)} \beta P(c_x) \quad (2)$$

where β controls how much the effect reacts. In this case, the actual propagation score of c_y will be the sum of the score fractions received through his associated candidates in S. Intuitively, the more expertise scores a candidate accumulates, the more probable that he is expert. The above algorithm is very similar to PageRank [2]. However, PageRank is based on a mutual reinforcement between pages and the score is computed iteratively to convergence. But for expertise propagation the candidates are only influenced by those ones in the seed.

3. Building associations among candidates

As we stated in the introduction, we hope to find the relationships between the candidates both from the web pages and email messages. We represent the organization whose corpora we study as a graph. The nodes correspond to candidates and the edges correspond to the strength of the associations among the people. In web pages persons co-occur in some local context may indicate that they are related under some topic while the existence of email correspondence also denotes the potential association among the people.

Through the analysis of the graph we can quantify the strength of the association between two people. A higher strength indicates that the two people have more common interest and more frequent communication.

3.1. Web pages-based Social network

To build the social network from web pages, we assume that people co-occur in a range of local context may share similar interest. Therefore an intuitive way of quantifying the strength of the association between two candidates is to count their co-occurrence in a document. We formalize the notion of co-occurrence of two candidates c_x and c_y in a document *d* by a binary function CO over all documents in the web pages collection:

$$CO(c_y, c_x, d) = \begin{cases} 0 & \text{if } c_y, c_x \text{ do not co-occur in } d \\ 1 & \text{if } c_y, c_x \text{ co-occur in } d \end{cases} \quad (3)$$

And the association between candidates c_x and c_y is obtained by:

$$a(c_y, c_x) = \sum_d CO(c_y, c_x, d) \quad (4)$$

However, the co-occurrence function defined above is not delicate enough to represent the association precisely. Since there may be many topics within one document, the further the distance between the two candidates is the less probable that they are under the same topic. This observation suggests that we weaken the strength of the association as candidates separate away from each other. To achieve this we describe the possible scheme that employing the reciprocal of the number of words between the occurrences as co-occurrence strength.

3.2. Email communication-based Social network

We propose several strategies to build the association among candidates through the analysis of email communication patterns.

The simplest way of estimating the strength of connection between two candidates is to count the amount of their email correspondence. The candidates c_x and c_y are associated if they appear together in the *from*, *to* or *cc* field of an email *e*. We introduce a binary function EC(c_x , c_y ,e) (Email Connection) to represent their connection:

$$EC(c_y, c_x, e) = \begin{cases} 0 & \text{if } c_y, c_x \text{ do not appear in } from, to, cc \\ & \quad \text{fields of } e \\ 1 & \text{if } c_y, c_x \text{ appear in } from, to, cc \text{ fields of } e \end{cases} \quad (5)$$

Email messages form a thread tree structure. The root of the thread tree is the first message sent by some candidate and then the thread tree expands as other people reply to this message or forward to others to continue the discussion. Since more candidates are involved in email threads, consider associating candidates appearing in the same email thread can tackle the problem that sometimes the email connection matrix built merely through single message is sparse. Similarly, we use binary function

TC(c_x , c_y ,t) (Thread Connection) to represent their connection:

$$TC(c_y, c_x, t) = \begin{cases} 0 & \text{if } c_y, c_x \text{ do not appear in } from, to, cc \\ & \text{fields of } t \\ 1 & \text{if } c_y, c_x \text{ appear in } from, to, cc \text{ fields of } t \end{cases} \quad (6)$$

In email collection the association between candidates c_x and c_y can be estimated by combining both resources, which is achieved by a linear interpolation smoothing over single message and its corresponding thread together.

$$a(c_y, c_x) = \sum_e ((1-\alpha) EC(c_y, c_x, e) + \alpha TC(c_y, c_x, t_e)) \quad (7)$$

where t_e is the email thread that e belongs to; α is the interpolation factor, which can be viewed as mixture weight if the equation is considered as a two-component mixture model.

3.3. Query dependent social network

The graphs built above reflect the static relationship among candidates, which is query independent. However, with the query topics involved we can build a query *dependent* social network among the candidates. In this situation, the documents adopted to generate social network are not from the whole collection. In stead we calculate the associations from those web pages and emails relevant to the query topic. This query dependent social network is appealing because it focuses merely on the associations which are related to the desired topic. For example, if one candidate is the expert in different realms including the desired one, the candidates to whom we want to assign expertise score are those communicating much with him on the required topic.

Moreover, we could better evaluate the strength of association by employing the similarity of the documents to the query. This is based on the following assumption: the more relevant to query the document that joins the candidates is, the stronger association exists among the candidates. Thus we replace the binary function with CO EC and TC with numerical function in which the weight is the similarity of the document to the query.

4. Experiment and discussion

Our CDD-based search model is evaluated on the dataset adopted in TREC enterprise track 2005 and 2006[3]. The collection is a crawl of the public W3C (*.w3c.org) sites, including Web pages and Email lists. We took the topics adopted by the expert finding task of TREC. The main evaluation measures used for the expert finding task is mean average precision (MAP).

4.1. Experimental results

The baseline of our experiment is the best results in the stage 1 as described in [1]. For the expertise propagation process, there are some free parameters to be estimated, for instance, the size of the seed N. In this part, we empirically tune the parameters to the best on the training topics and then test on the rest topics. Table 1 shows the results the expertise propagation under different kinds of social network, including co-occurrence association CO and communication association EC and TC.

Table 1: Comparison between different kinds of social network

Social Network	**TREC 2005**		**TREC 2006**	
	MAP	P@20	MAP	p@20
Baseline	0.2847	0.3469	0.4592	0.4854
CO	0.2882	0.3518	0.4621	0.4955
DW CO	0.2998	0.3617	0.4700	0.5070
EC+TC	**0.3101**	**0.3701**	**0.5075**	**0.5292**
QD EC+TC	0.2995	0.3600	0.4884	0.5071

We see that the co-occurrence association outperforms the baseline over the two topics set, which indicates that CO is helpful in identifying that persons often co-occur in the context as a group. And the performance gap between the association CO in equation 4 and the distance weighted CO (DW CO) confirms our assumption that there are multi topics in one document and associating candidates who are far from each other may cause the topic drift problem.

The expertise propagation using social network built through analysis of email communication also performs well. The reason lies in the fact as stated in [4] that Chen et al calculate the clustering coefficient [5] of the email network in W3C. The clustering coefficient is 0.267, which is much larger than the clustering coefficient of the comparative random network where coefficient is 0.00041. It indicates that the mail network is highly clustered and our experiment approves the fact. However, we can observe that the improvement of query dependent social network is not as significant as the query independent one. We will discuss this in section 4.3.

4.2. The role of the seed

In expertise propagation algorithm, the candidates in the set of seed will assign expertise probability to other candidates who are related to them. Thus one crucial parameter in expertise propagation process is how many candidates would be selected as the seed set. Varying the number of candidates allow us to evaluate the effect of the seed on the test sets. The curves shown in Figure 1 reflect the variation caused by number N.

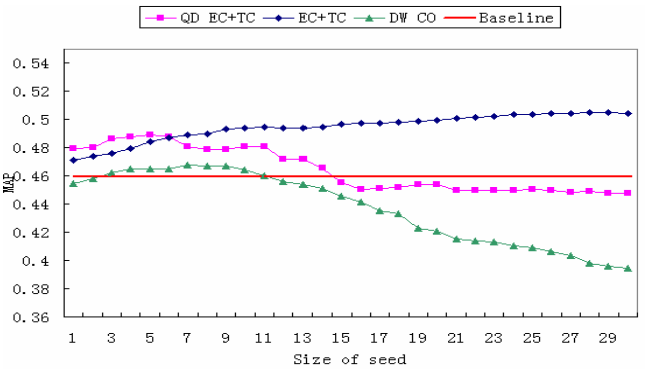

Figure 1. The impact of varying the size of seed in tuning the effect of expertise propagation on set TREC 2006.

When the number is limited to a small range, the size of the seed is not large enough to influence many candidates and this may cause a bias problem that only favoring people associated to the few high-ranking candidates. With the size enlarged, too much noise may be covered within the seeds that some candidates having high expertise probabilities are not truly relevant to the query. However, the query dependent social network outperforms the query independent ones for size smaller than 8 then its effect decrease. These results contradict to our assumption a little bit that the query dependent social network would help reduce the influence of the noise since it filters those messages irrelevant to the query. The results show that the query dependent social network in email and co-occurrence social network are sensitive to the size of the seed. This can be explained by the fact that both of the social networks built are sparse: among 46059 web pages, total 8100 of which contains more than two candidates while out of 198394 emails in the collection only 17106 emails are relevant to the topic on average. The sparser network may not represent the truly distribution of the association and may cause the small sample problem [6].

4.3. The role of weights estimation

The intensity of the expertise propagation is controlled by the weight parameter β in equation 2. In our experiments, we do not use training methods to optimize the weights automatically, in stead we tune it empirically on two topic sets to reveal and validate the impact of the intensity. We vary the value of the parameter β and record the performance at each parameter value. The results show that the insufficient intensity is more harmful than over emphasizing the expertise propagation effect. How to keep balance between the two aspects is crucial and the empirical results show that setting β from 0.1 to 0.3 is beneficial.

Another parameter needs examination is the interpolation factor α in equation 7 which keeps balance between the thread connection and the email connection in building association through analysis of communication patterns. We find that as the α increases the performance is enhanced until α is close to 1, which suggests that the thread connection effect should be highly emphasized to optimize the performance. This phenomenon may be still due to the sparse data problem as stated in the previous section that the email threads associate more candidates than email messages do.

5. Conclusion and future work

In this work, we introduce an expertise propagation algorithm to adjust the expertise relationship between candidates. The social networks are built from two resources: on one hand, people explicitly send email to one another thus the relationship are likely to be contained in the patterns of communication; on the other hand, we can rely on statistical relationships extracted from co-occurrences of people in web pages. The performance improvement in our experiments demonstrates its effectiveness. However, while the expertise propagation performs well with the actual experts in the seeds, it is not robust to the noise and some candidates in the seed are not truly relevant to the query.

In the future we will be dedicated to tackling the problem of reducing the influence of the noises in the seed to the expertise propagation. We also plan to investigate other kinds of potential association among people and to combine the current associations into a general social network.

6. References

[1] Y. Fu, R. Xiang, M. Zhang, Y. Liu, S. Ma, A PDD-based Searching Approeach for Expert Finding In Intranet Information Management. *The third Asia Information Retrieval Symposium, AIRS06*, 2006,

[2] L. Page, S. Brin, R. Motwani, T. Winograd, *The pagerank citation ranking:Bringing order to the web*. Tech. Rep. Computer Systems Laboratory, Stanford University,Stanford, CA. 1998.

[3] TREC. Enterprise track, 2005 and 2006. URL: http://www.ins.cwi.nl/projects/trec-ent/wiki/.

[4] H. Chen, H. Shen, J. Xiong, S. Tan, X. Cheng, Social Network Structure behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track. *The Fourteenth Text REtrieval Conference Proceedings (TREC 2006)* , 2006.

[5] M.E.J. Newman, The structure and function of complex networks. *SIAM Review,* 45(2):167–256, 2003.

[6] N Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 2002 - IOS Press