Pitt at TREC 2006: Identifying Experts via Email Discussions

Daqing He, Yefei Peng
School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260
{dah44, yep3}@pitt.edu

Abstract: Identifying experts in a certain domain or a subject area has always been a challenge in various settings including commercial, academia, and governmental institutions. Our interests in this year's TREC Enterprise track are to utilize the email communications as the basis for identifying experts and their expertise on certain topics. In this report, we presented a method for identifying experts based on the emails they sent around. We hypothesize that experts would be more active in relevant email threads, would send longer emails, and would participate in the discussion at the very beginning of the threads. An algorithm based on these hypotheses was developed and tested in this year TREC enterprise track experiments to find experts for 49 topics based on documents in the W3C collections. Our initial experiment results produced suboptimal performance. This motivated us to examine the hypotheses more closely in the context of provided ground truth. Interestingly, the analysis on ground truth seems to confirm that all of our hypotheses have their merits in finding experts, so one future important question is how to utilize these rules in a right way.

1 Introduction

Identifying experts in a certain domain or a subject area has always been a challenge in various settings including commercial, academia, and governmental institutions. The information about the where about of certain expertise has been maintained by social connections and personal relationships. However, two developments in the past several years have significantly changed the landscape of the expertise management. The first is that many organizations have accumulated vast amount of digital documents produced by their employees on various tasks in multiple media (especially emails), and the documents are often stored and organized in organization premises, which makes the access of those materials much more easier than before. The second development are the rapid development of information processing techniques, especially in the areas of natural language processing and information retrieval (Balog et al. 2006; Craswell et al. 2005).

This year's Enterprise Search Track represents the continuous effort of developing new information processing techniques to handle the expert search problem in the context of identifying potential experts by examining various forms of digital documents, such as emails, web pages, personal homepages, etc (Soboroff et al. 2006). Utilizing the W3C collection[1] crawled at the public W3C (*.w3.org) sites in June 2004, the collection consists of emails, web pages, memos, etc., which provides a realistic data set for finding experts on certain topics within an organization, which in this case is W3C.

Our interests in this year's TREC Enterprise track were to utilize the email communications as the basis for identifying experts and their expertise on certain topics. We think that W3C emails represent certain type of discussion exchanges among people who share the same interests on certain topics. This may cause some aspects of the emails in W3C collection to be different to those emails in more casual settings. Therefore, our goal in this year's study was to find a set of useful features from the W3C email exchanges that could be used as evidence for identifying experts. We paid specific attention to features like the threads, the

[1] http://research.microsoft.com/users/nickcr/w3c-summary.html

position of an email in a thread, the length of the first email in a thread, and others that can be utilized as heuristics for designing algorithms.

In the remaining of this report, we will first talk about our approach in detail, then present the experiment design including the runs that we submitted. We also will present our result analysis including the study of ground truth to examine the merits of the hypotheses that were used in our approach. We will conclude with some final remarks and future directions.

2 Our Approach

We model the expert search as a process of finding a set of candidates who have potential knowledge of a given topic that is expressed as a query. We concentrated on emails because email documents seem particularly well suited to expert search as people routinely communicate what they know (Campbell et al. 2003)

Through observing some message exchanges in W3C email collection, we hypotheses that the experts on a given subject area might have the following characteristics:

- *Hypothesis 1*: *experts are more active in the threads that are related to a particular topic*. This seems to be intuitive, and is the underneath motivation for judging the expertise of a person based on how many relevant emails he/she has sent on a subject area.
- *Hypothesis 2*: *experts usually send long emails*. W3C email collection contains not normal emails. It stores emails of some discussion groups. Although we can see experts sending short replies sometimes, lots of time what we see are experts sending drafts of protocols, standards, an so on. All of these emails are rather long documents.
- *Hypothesis 1*: *experts usually start to participate in the discussions at the very beginning of the threads*. We find that threads in W3C email collection often start with two types of scenarios. A thread may start with a person sent out a long document about certain protocol or standard for other people to comment on. We think that the person is an expert on that subject area since he/she played a leading role of writing up a relevant document and initiating a discussion. A thread can also start with a person sent a message to one or a set of person to ask for help or to clarify a question on a topic. In this case, the people being addressed here are the experts otherwise they would be not asked for help. No matter under which scenario, the person who is an expert involves in the discussion at the very beginning.

Of course we acknowledge that the above hypotheses were obtained purely by our observations in the collection, which has not been tested by experiments. However, they do provide us some form of potential heuristic rules for differentiating among emails and among the persons who sent those emails.

Based on the above observation, our expert search process consists of the following steps:

1. *Initial Search*. Since all our hypotheses rely on a set of relevant email to a subject area as the start, a search engine is used to locate a set of potential relevant emails. The follow-up steps will concentrate on identifying experts based on the returned emails.
2. *Score Calculation*. some form of algorithm based on the above three hypotheses is used to analyze the emails and rank potential expert candidates who are associated with the emails. Here we assume that the people who receive higher score would have higher possibility to be the experts.
3. *Candidate Ranking*. candidates that have non-zero scores are ranked to generate the final result.

The first step will use a standard retrieval engine, whose details will be presented in Section 3 about the experiment design. Here, the remaining part of this section will present our algorithm for ranking emails based on their characteristics and their positions in the thread, then we will present the algorithm for ranking people based on the emails that they associated with, which eventually give us the clues who could be the experts.

As a summary of the emails that could lead us to experts, we prefer an email that is long in its true length. Here the true length is the real new content in an email that excludes the cited part from previous emails. We also prefer the first message in a relevant thread. However, we do notice that short first message in a thread usually corresponds to the scenario that an non-expert asked some questions, so we want to further specify that we prefer the first message that is long, and decrease the weights for a short first message in a relevant thread. Finally, we prefer emails that are at the beginning of the threads and the deeper an email is in a thread, the less weight we assign to it. Therefore, suppose N_i is the total number of emails in thread *i*, *T* is a length threshold that differentiate a long first message to a short one, then the weight $W_{i,j}^T$ for an email at *j*th position in thread *i* can be defined as

$$W_{i,1}^T = \frac{N_i - j + 1}{N_i} \times 0.5 \text{ , if } L_{i,1} < T$$

$$W_{i,1}^T = \frac{N_i - j + 1}{N_i} \times 1.5 \text{ , if } L_{i,1} > T \quad (1)$$

$$W_{i,j}^T = \frac{N_i - j + 1}{N_i} \text{ , } j > 1$$

These formulas give relative low weight (e.g. 0.5) to those first messages that are short (< T), but give much higher weight (e.g. 1.5) to those first message that are longer than the threshold. The weights of non-first messages are proportionally reduced along with the increase of the position of the emails in the thread.

Then based on the hypothesis that experts are those active in relevant threads, we take a simple evidence combination approach. We use the sum of the scores of all the emails in the result list that are related to one candidate to calculate the final score for predicting the expertise of the candidate. Therefore, the score S_k for candidate *k* is calculated as

$$S_k = [\sum_{i \in AllThread} \frac{\sum_{j \in S(i,k)} W_{i,j}^T W_{i,j}^I L_{i,j}}{\sum W_{i,j}^T W_{i,j}^I L_{i,j}}] \log \frac{|T|}{|T_k|} \quad (2)$$

Where

$W_{i,j}^T$ is weight in thread for jth email in ith thread, and it is calculated in (1)

$W_{i,j}^I$ is weight from the initial email retrieval for jth email in ith thread. The weight is generated by the search engine.

$L_{i,j}$ is true length of jth email in ith thread.

S(i,k) is emails from candidate k in ith thread.

|T| is total thread number

|Tk| is number of thread in which candidate k shows up.

3 Experiment Settings

The expert search task of Enterprise track used W3C collection, whose email sub-collection contains 198,394 emails, which was the sole document evidence we used in finding experts. The test collection also has 49 topics that ranging from asking for experts on choreographies in semantic web to those on authoring tools web accessibility guidelines.

We used Indri 2.0 as our initial search engine, and used the thread information produced by William Webber during TREC 2005 Enterprise track. In total, we submitted 4 different runs. They are:

- **PITTNOPH**: In the initial Indri query, both the title and the description were used, and each word was treated as a term in the query. The full formula (2) was used to calculate the score for expert candidates.
- **PITTPHFULL**: In the initial indri query, all the words in the title part were treated as a phrase, but each word in the description was treated as individual term.
- **PITTPHFREQ**: In the initial indri query, the title and the description were treated the same as in PITTPHFULL. However, the calculation of expert scores was done by a simplified version, which only look at the sum of the Indir scores of the returned emails.

$$S_k = [\sum_{i \in AllThread} \sum_{j \in S(i,k)} W_{i,j}^I] \log \frac{|T|}{|T_k|} \quad (3)$$

- **PITTMANUAL**: using the algorithm of PITTPHFREQ, we generated 20 candidates for each query, and each candidate has up to 5 support documents. We then had three people manually select expert candidates based on the support documents.

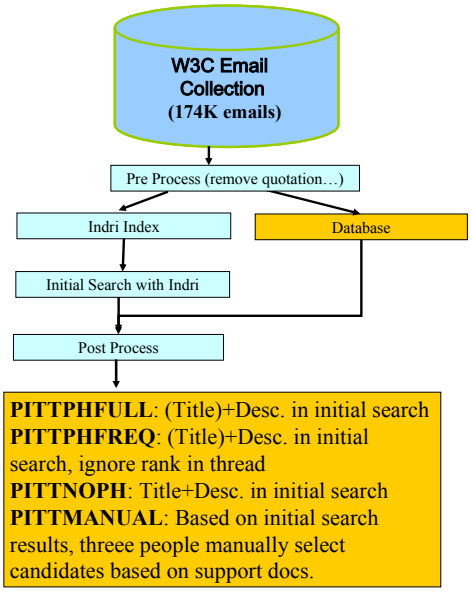

Figure 1: The Experiment Settings

4 Results Analysis and Discussion

4.1 Analysis of Submitted Results

As shown in Figure 2, overall our elaborated approach PITTPHFULL, which identifies experts based on formula (2) did not receive much benefits. In actual fact, it performs worse than the simpler approach that counts only the emails that were associated with a person and the relevance of these emails (i.e., PITTPHFREQ). This inferior result appeared in all four measures (see Figure 2), although the difference was not significant.

Our manual effort for identifying experts (i.e., PITTMANUAL) did not pay off too. It actually performed the worst in three out of four measures. Only when measured by precision at 10 (P@10), it is the second best run. This demonstrates one problem with our manual approach. Although human involvement can help at identifying experts more accurately at the top of the candidate list, due to the large quantity of topics, candidates in each topic, and the supporting emails for each candidate, human could not go through every email of every candidate for every topic, and thus the recall is dropped as the consequence.

We can think of two major reasons why PITTPHFULL did not perform better than PITTPHFREQ. The first one is that we did not develop a good implementation of the ideas, whereas the other is that the hypotheses are not true, which trigger the failure. Since the second reason is much more serious, we will report our investigations of them in detail in the remaining of this report.

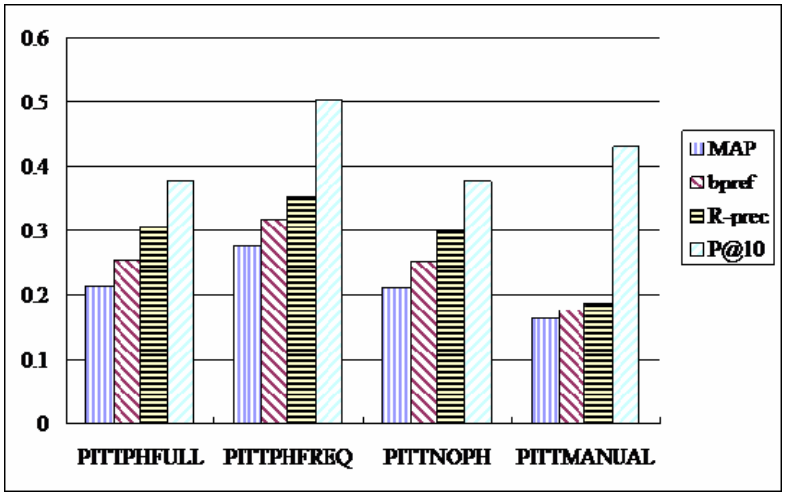

Figure 2: The results of the four submissions

4.2 Analysis of the Hypotheses in Ground Truth

Our analysis of the hypotheses was performed on the ground truth directly. Although we acknowledge that the ground truth may not be perfect due to the distributed annotation process employed in the Enterprise track, a close examination of the characteristics of the expert lists in ground truth and their corresponding support emails may still give us some insights.

4.2.1 Hypothesis 1

Our first hypothesis states that experts are more active in the relevant threads to a particular topic. As we mentioned earlier, this is intuitive and important since we basically use such information as the motivation for combining evidence to find experts. Our analysis of the ground truth examined the number of relevant emails from experts that are in the relevant threads, and then made comparison to the number of emails from a person who we do not differentiate whether he or she is an expert. On average experts send 8.3 emails in the relevant threads, and among them 4.8 emails are support emails, whereas on average a person (i.e., including both experts and non-experts) sends only 2.4 emails in relevant threads. It seems that experts indeed are more active than an average person in the sense of number of emails sent.

We also examined a variant of hypothesis 1. It states that "*experts often involve in more numbers of relevant threads than non-experts*." The analysis is done at counting how many relevant threads an expert and a non-expert involved in.

Figure 3 shows the result of this analysis. On the x-axle we get the number of relevant threads from 1 to 25, and on they-axle are the number of experts or non-experts. It is clear that the distribution of experts related to the number of relevant treads is dramatically different to that of non-experts. On average, each expert appears in 4.5 relevant threads, whereas each non-expert appears only in 1.8 relevant threads.

Therefore, it seems that experts and non-expert not only have different numbers of their emails in relevant threads, but also have different numbers of relevant threads they involved in.

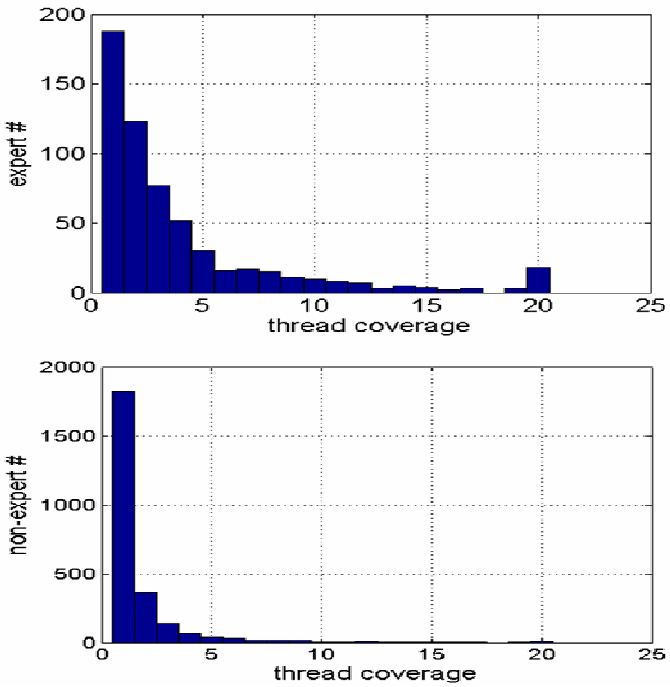

Figure 3: The activity comparison between experts and non-experts

4.2.2 Hypothesis 2

The second hypothesis states that experts usually send long emails. We examined this hypothesis by looking at the true length of relevant emails used as the evidence to confirm the expertise of a person (i.e., called support emails). If this is greatly different to the average length of the emails in the collection, we probably can think that the hypothesis has some truth there.

The hypothesis seems to be confirmed by the ground truth data. The average length of the support emails is 544 words, whereas the average length of emails in the whole collection is only198 words.

We again examined a variant of hypothesis 2. We hypothesized that "*the first email sent by an expert is usually a long email*." According to the ground truth, the average length of the first messages sent by experts is 690 words, whereas that of all first messages in the whole collection is only 211 words.

Therefore, it seems that experts and non-experts not only sent emails with different true length, but they sent the first message with different truth length too.

4.2.3 Hypothesis 3

Hypothesis 3 states that experts usually start to participate in the discussions at the very beginning of the threads. Our analysis therefore looked at the positions of the support emails. Here the position is relative to a thread, where *position n* means that the email is *nth* message in the thread.

As shown in Figure 4, the average position of experts' support emails is 3.6, whereas more than 50% of relevant messages are the first message in the threads, and 80% of support messages to experts are among the first four messages in the threads.

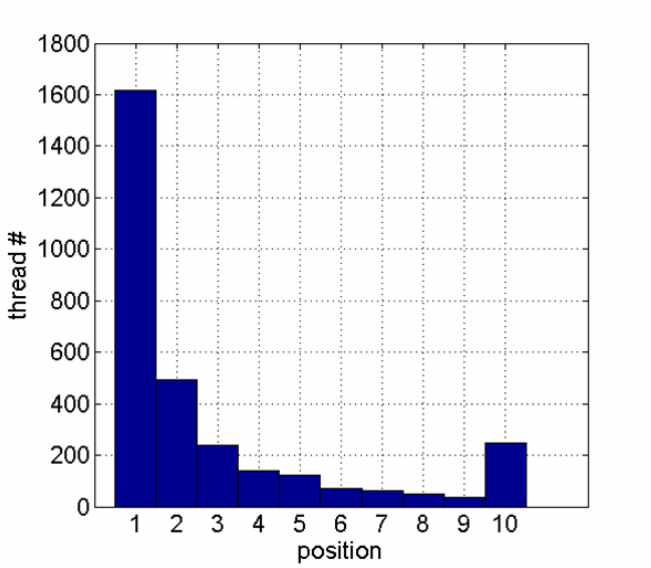

Figure 4: The positions of relevant support emails in the threads

Therefore, it seems that experts indeed participate in the discussion at very beginning of the threads.

5 Conclusion

In this paper, we presented a method for identifying experts on some subject areas based on the emails they sent around. This was done as part of our participation to the TREC Enterprise Track 2006. We hypothesized that experts would be more active in relevant email threads, would send longer emails, and would participate in the discussion at the very beginning of the threads. We therefore built an algorithm based on these hypotheses to identify experts. The algorithm was tested in the experiment of finding experts for 49 topics in the W3C collections. Our initial experiments generated poor performance results. As part of failure analysis, we examined the hypotheses more closely in the context of provided ground truth. This part of analysis seems to confirm that all of our hypotheses have their merits in finding experts, so the poor performance was probably due to the fact that we have not found the right way to utilize them.

Our future study will be concentrated on further analysis of the performance results, and develop a better method to integrate the confirmed hypotheses into the expert search algorithm.

6 Acknowledgement

Thanks Ian Soboroff, Arjen de Vries and Nick Craswell for organizing the Enterprise track. This work has been supported in part by DARPA contract HR0011-06-2-0001.

7 Reference

Balog, K., L. Azzopardi & M. d. Rijke, (2006). Formal Models for Expert Finding in Enterprise Corpora, in *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Seattle, WA, USA, 43-50.

Campbell, C. S., P. P. Maglio, A. Cozzi & B. Dom, (2003). Expertise Identification using Email Communications, in *Proceedings of Conference on Information and Knowledge Management* New Orleans, LA: ACM.

Craswell, N., A. de Vries & I. Soboroff, (2005). Overview of the TREC 2005 Enterprise Track, in *Proceedings of TREC 2005*.

Soboroff, I., A. P. de Vries & N. Craswell, (2006). Overview of the TREC 2006 Enterprise Track, in *Proceedings of TREC 2006 Conference* Gaithersburg MD USA.