CNDS Expert Finding System for TREC2005

Conglei Yao Bo Peng Jing He Zhifeng Yang
School of Electronics Engineering and Computer Science
Peking University
{ycl, pb, hj, yzf}@net.pku.edu.cn

Abstract

This paper describes our system developed for Expert Finding task of Enterprise Track for TREC2005. This system employs 3 methods, traditional IR method, email clustering method and entry page finding method, to find experts related to a specific topic in W3C corpus. Experiment indicates that traditional IR method is useful to expert finding if the query is well generated, email clustering method is helpful when the mail list is relevant to a unique work group or committee, and entry page finding method is valuable while the topic is the theme of a special group. We use result aggregation methods of linear synthesis to combine the results generated by the three methods,. Of our 5 runs submitted for Expert Finding task, the best run is the one generated by linear synthesis, providing a MAP score of 0.2174(Bpref of 0.4299 and p@10 of 0.3460)

Keywords

Expert Finding, IR, Email Clustering, Entry Page Finding, Result Aggregation

1. Introduction

Expert Finding task is one of the two tasks of Enterprise Track this year. The goal of this task is to retrieve a list of candidate experts with each a rank score to a specific topic, in other words, to find all the related experts whose daily work is closed to the topic, and rank them according to the relevant extent .

To find the experts for every topic, we employ three methods based on the characteristics of the full W3C corpus and the topics. First, we use traditional IR techniques to create the query based on the topic content, retrieve the most relevant web pages, get the embedded person name and return them, the key issue of this method is generate the best query which can present the information need of expert finding task. Second, we analyze the main feature of mail list archive, cluster the senders' name based on the relationship of sending, receiving or replying, establish the link between the clusters and topics, and then get a list of candidate experts for every topic. Third, we extract the anchor texts and title for every page, and use them to get the entry page for every topic which is the main theme of a work group; meanwhile, we make use of them to get personal page for every person; then based on the

entry pages of work groups and personal web pages, we extract the relationship between persons and topics, and produce a list of relevant expert id for every topic. Finally, we use two result aggregation methods of linear synthesis and Markov chain to combining the three results generated by the former methods.

2. Methodology

2.1 Traditional IR techniques

During indexing step, first of all, we do some further preprocessing on the corpus. we remove stop words, stem with K algorithm, and then we get a clean corpus. Then we can build the index upon such clean corpus. We build the index with the term position information.

In Retrieval module, We generate the queries not only from the topics but also from the candidate person name, so we can evaluate something more than only “appearance” information. For query generation, we firstly need to generate query from topic and person name. We consider three kind of queries such as Boolean query, text query, and some other structured query (as lemur supplied), and select the Boolean query according to the experimental results.

We initially construct the query as AND operation of the topic phrase and person name phrase. If we can get some results from the retrieval system, then it stopped. Otherwise, we will iterate to transform the query to a suitable one. Such transforming includes confusing the phrase operation, cutting some terms with less information and so on. The terminal condition is to get a suitable number of retrieved documents.

Finally, we begin to ranking the results. We get the subset of documents which are more or less relative to the topic and the expert name, then we calculate a score to scale how relative they are. There are many approaches to represent the similarity between document and query. In this Task, the final document score is the combination of cosine value and the span based metric value.

Then we will calculate the score of a expert name. The evidences to prove a person to be a expert are the documents number and the score of the document. There are also some choices to calculate the final score, one is to get the highest document score as the person score, another way only consider the number of documents retrieved , and other one can consider the number of documents whose score is higher than a threshold. We mixed these approaches.

2.2 Email clustering method

First, we analyze the mail list pages, and extract information such as author, author’s e-mail address, mail list the mail belongs to. Then we group them by mail list name and count the number of mail sent by a same author, and get the relationship:

<mail-list, mail-address, doc-count>

The problem now is to label a topic for each mail list. We get it in three methods and combine them together. We use each topic to query the IR engine and count how many documents belong to each mail list. The more

documents there is, the more possible the mail list is relative to the topic. The second method is counting how many words are matched in the mail list's name and topic's description. The third method is based on entry page finding method to find the group's homepage. Once we find a group's homepage, usually we can find their mail lists mentioned somewhere in the page. So we get the third score to judge if a mail list belongs to a topic. Finally, we combine the scores together and figure out the following relationship:

<topic, mail-list, mail-list-score>

Now given a topic, we can find some mail lists which are relative and the corresponding mail addresses which have sent mails to the mail lists. Mail addresses are identical with candidates' IDs. So we get the score for each candidate of each topic by multiply the corresponding mail-list-score by doc-count. Note that there are people send mail to different mail list which belongs to the same topic, so we need to merge together the scores of the same candidate for a topic.

2.3 Entry page finding method

It's a fact that there are many work groups whose work focused on a special topic in an enterprise or organization have their entry web pages to show information, and many employees have their own personal web pages to show their interest and working advance. We analyze the training set and discover a truth that a large part of topics are related to work groups, and all the 17 employees' personal web pages we got manually describe the names and work direction of them. As a result, if we can find the entry page of every topic related to a specific work group, and get the personal homepage for every employee, we can precisely extract the relationship between topics and persons in the enterprise or organization.

Some researchers have studied the performance of different features which can be used to find entry page, and have concluded that URL form is especially useful. But for a relative small corpus such as the full w3c corpus, because of the different goal of creating web pages (just for revealing information but not for attracting visitors' attention), the URL form is not the same as the web pages of the large corpus of Web. Meanwhile, the link structure and the characteristics of anchor texts are different from the the large Web. Therefore, in order to find the entry page, a new methodology must be proposed to resolve this problem, we analysis the influence of web pages' indegree, anchor text, URL form and title to entry page finding, and create a new algorithm of finding entry page which integrates the former 4 features. Experiments have indicated the effectiveness of this algorithm. For the discovery of personal web pages, we just use the personal name and email as the query to find entry page of some person, and get useful results. When all the entry pages and personal homepages have been found, analyzing their content, we can extract the relationship between topics and persons.

3. Conclusion

Expert finding task based on a corpus of an enterprise or organization is different from the traditional search problem resolved by popular search engines, not only because of the new characteristics of data, but also because of the new feature of information need embedded in this task. To resolve this problem, some other methods based on the corpus should be considered to combine with the traditional IR methods. This paper employs two methods of mail

archive analysis and entry page finding to work with traditional IR methods, and the result indicates the effectiveness of this methodology.

Acknowledgment

This work described therein is supported in part by PRC Ministry of Education grant 20030001076 and by NSFC grant 60435020.

References

[1] Javier Artiles, Julio Gonzalo, and Felisa verdejo. A Testbed for People Searching Strategies in the WWW. *SIGIR'05, pages 569-570,* 2005.

[2] Nick Craswell, David Hawking, and Stephen Robertson. Effective Site Finding Using Link Anchor Information. *SIGIR'01, pages 250-257*, 2001.

[3] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma, Hong-Jiang Zhang, and Chao-Jun Lu. Implicit Link Analysis for Small Web Search. *SIGIR'03, pages 56-63,* 2003.

[4] N. Eiron, and K. McCurley. Analysis of anchor text for web search. SIGIR'03, pages 459-467, 2003.

[5] Ronald Fagin, Ravi Kumar, Kevin S. McCurley, Jasmine Novak, D. Sivakumar, John A. Tomlin, and David P. Williamson. Searching the workplace web. *WWW2003, pages 366-375*, 2003.

[6] Hamish Cunningham, Yorick Wilks, and Robert J. Gaizauskas. Gate: a general architecture for text engineering. In *Proceedings of the 16th conference on Computational linguistics, pages 1057-1060, Morristown, NJ, USA*, 1996. Association for Computational Linguistics.

[7] P. Ogilvie and J. Callan. Experiments using the lemur toolkit. In *Proceedings of the 2001 TREC conference*, 2002.

[8] R. Krovetz. Viewing Morphology as an Inference Process,. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 191-203,*1993.