THUIR at TREC2007: Enterprise Track[1]

Yupeng Fu, Yufei Xue, Tong Zhu, Yiqun Liu, Min Zhang, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Yupeng.Fu@gmail.com

Abstract. We participate in document search and expert search of Enterprise Track in TREC2007. The motive behind the TREC Enterprise Track is to study the issues searching the documents and experts inside an enterprise environment, which has not been sufficiently addressed in research. In document search, we focus on the key overview page pre-selection methods and link analysis algorithms. In expert search, we develop methods to detect expert identifiers and experimented based on our previous PDD model.

1 Introduction

This is the third year that the IR groups of Tsinghua University participated in TREC Enterprise Track. Different from previous tracks, TREC introduced a new enterprise corpus and new tasks. The approaches we've studied this year include link analysis among documents, person entity identification, topic distillation with key resource pre-selection, results combination and some other technologies.

For document search task, we mainly investigate the effects of key source pre-selection and link analysis among the documents. We first observe the high quality resource distribution. Some features are studied to find overview pages. We also do some link analysis: both HITS and PageRank algorithms are employed to evaluate the page quality. Besides, we attempted a novel link analysis method which involved the document similarity.

For expert finding task, a lot of efforts have been made on name identification. We built personal description documents (PDD) for each candidate from various types of resources. We obtain retrieved results from each description document collection. And with the help of EM algorithm we combine the results from different corpus to generate a merged ranking list.

2 Document Search

The task of document search is defined to retrieve those documents which help the science communicator create an overview page in the given topic area. These will tend to be authoritative pages such as project homepages and documents dedicated to the topic, rather than pages that make passing mention of the topic. There are two potential approaches to find those “overview page”. The first one is to build a query independent classifier that selects those documents with specific features to be required key pages. The other one is to adapt link analysis to predict those authoritative pages. Both approaches were attempted in our experiments.

2.1 Key pages pre-selection using query independent features

To retrieval key pages, we can first retrieval query relevant pages from the whole corpus then mine those overview pages. However, the volume of the whole corpus is large that may take much time to retrieve. Therefore we try to do some data cleansing work to pick those key pages out according to some query independent features. From the example pages provided by NIST and some other overview pages browsed we noticed that the overview pages are under similar templates: many out-links, many internal links and similar design of layout. However, we believe that recognizing overview exactly according to the template of layout is not robust. So we tried to build a classifier to select overview pages using features like amount of out-links and

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108)，Natural Science Foundation (60621062, 60503064，60736044) and National 863 High Technology Project (2006AA01Z141)

in-links.

Figure 1 shows that the distribution of the amount of out-links in overview pages provided by TREC. From the distribution we notice that most overview pages contain a lot of out-links from 115 to 180, contrast to the fact that among the whole corpus more than ninety percent of pages are with out-links less than 100. We conduct an experiment that only retrieve from those documents which have number of out-links more than 100. We use the example overview pages as relevant pages. The results show that the performance is 10% higher than retrieving from the whole corpus while the size of the selected corpus is only 10% of the whole one.

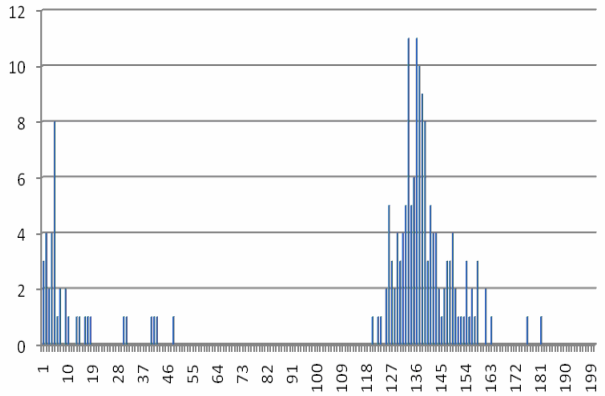

Figure 1. The distribution of the amount of out-links in example pages

Similarly, we make the hypothesis that the overview pages are those authoritative ones which may be linked by hub pages. Therefore the in-link anchor text may provide much useful navigational information. The experiments results validate our supposition that the anchor text based retrieval outperform the full text retrieval by more than 20% in MAP.

Some other features are also attempted to identify overview pages, such the length of page URLs, the amount of in-link. We also implement a decision tree to combine these features to better construct the overview page corpus. Finally, the performance of the results retrieved from the document set after distillation achieved 35% improvement than from the full text collection.

2.2 Adapting link analysis for finding authoritative pages

While authority of pages on the Internet has been an active area of research, estimating authority of web pages inside an enterprise such as CSIRO is an open question. Since the difference of the amount of web pages in corpus and the link structure, whether the previous link analysis algorithms such as HITS and Page Rank is effective is need to explore.

We first used Page Rank [1] to estimate the quality of the web pages. However, the link structure of the web site CSIRO is quite different from the environment that the algorithm applies to. So after executing the Page Rank process, we get some pages with very large scores like the homepage and index pages. However, it is difficult to use these importance scores in conjunction with query-specific IR scores to rank the query results. Several known approaches are attempted but all failed to improve the performance.

Another famous link analysis algorithm we tried is HITS [2]. This idea has an intuitive parallel for finding overview pages. Although the pages with high authority scores are those overview pages, it’s hard to combine these scores with content based scores. We tried to use EM algorithms [3] to aggregate them through linear combination but the improvement is trivial.

The overview pages we want to find more or less are both good hubs and authoritative pages, which means the required overview pages link to important pages and other good pages also links to them. Therefore for a given page, we add the similarities of those pages which the link to the page as the new score. This reflects the idea that considering the similarity of pages instead of merely analyzing the link relation. For example, if a hub

points to an authoritative page and the hub page is a good one, then the authoritative may get a lot of reinforcement. However, if only a little part of the hub page is relevant to the page it links to, the authoritative page may get too much. Plus the strength of the link relation is more or less reflected by the anchor text. Therefore involving the similarity of anchor text to the algorithm may better quantitatively determine the strength of reinforcement. The experiments results show that the improvement achieved by the content-based link analysis is consistent and significant.

2.3 Evaluation results

In this year document search task, we used the examples pages as relevant results to train our systems. In total we submitted 4 runs which are listed in table 1.

Table1. The official results of document search runs

	MAP
THUDSANCHOR , only anchor text	0.1181
THUFULLSR, full text with SimRank	0.3427
THUDSSEL, Selected Sub-collection	0.1161
THUDSSELSR , Selected Sub-collection with SimRank	0.1347

However, the performance results are much out of our anticipation. The reasons lies in that first, the documents retrieved from sub-collection are much less than those retrieved from whole collection. So in the pooling process, there may be some documents our runs do not cover; second, the training set we used consists of example pages from topics, which may cause over-fit due to its small scale.

3 Expert Search

There are some differences from previous Expert Search Task. One main difference is that there is no master list of candidates. So our system should automatically detect and find expert identifiers, such as email address, names. We make a lot of efforts on person name entity recognition this year. Another important difference is that the judgments are more accurate than judgments made in previous years. The set of relevant experts is small because users only want to see the main contacts for their query, not a list of 10+ knowledgeable people. This setting quite makes sense. It requires us to design strategies to better rank candidates.

3.1 Expert identifiers detection

Because there is no master list of candidates, we should automatically detect expert identifiers. According to the guideline that among CSIRO the pattern of the emails is firstname.lastname@csiro.au, we extract all the email addresses from the corpus, including some variation such as that the @ symbol may be HTML-encoded. In total we get 3170 addresses. After eliminating the typo in name and name variations in addresses, we got 3131 candidate names from the email addresses.

There are some variations for English names. Given a name “firstname.lastname“, at least five variations are possible: Firstname Lastname, Firstname.Lastname, Firstname Middlename Lastname, F. Lastname, F. M. Lastname. For the first three pattern of variation, we use Aho-Corasick algorithm to label. It takes O(m+n) time, where m is the length of the name and n is the length of the document. Labeling F.Lastname is a problem that a bit hard to tackle. Because it is possible that more than one person shares one abbreviation. For example, T. Thomas may represent Tom Thomas or Tim Thomas. Among the 3131 candidates, there are 162 ambiguous abbreviations. We tried to eliminate the ambiguity according the co-occurrence of other labeled names. However, there are still some ambiguous abbreviations hard to eliminate.

Some other technologies such as pronouns eliminating are also integrated in our system.

3.2 Constructing PDD and merging results

As we did in previous expert finding task, we build person description document (PDD) for each candidate [4]. We extract some candidate relevant information from the document as expertise, for example, the context around the expert identifier.

Besides, we also extract information from candidate's homepages. In total we find 477 homepages, which are about 15.2% of the amount of candidates. We name the PDD constructed from the homepages as detailed PDD (DPDD). There are two other collections of PDDs we built. One is from the anchor text while the other is from the key overview pages corpus as described in document search section.

To merge the ranking results retrieved from each PDD collection, we tried EM algorithm to assign the weight to each ranking similarity. When the ratio of the weights is parallel to the ratio of the MAP achieved by each ranking list, the merged list achieves the best performance [5].

3.3 Evaluation results

In this year expert search task, we submitted 4 runs which are listed in table 2. The results show the effectiveness of our combination of PDDs.

Table2. The official results of expert search runs

	MAP
THUIRPDD2 , baseline	0.4300
THUIRPDD2C40, with the size of window 40	0.4499
THUIRMPDD2, linear combination of PDD1,2	0.4122
THUIRMPDD4 , combination from all 4 PDDs	0.4632

4 Discussion and Future Work

In document search task, we attempted a promising content-based link analysis algorithm. We find the link structure of intranet and a website is totally different from the Internet. Therefore we will try to investigate the link analysis algorithms applying in intranet. A bigger ambition is that we would like to experiment our content-based link analysis algorithm on bigger data sets and to propose an algorithm integrating the link analysis and similarity ranking.

For expert finding task, in the past two years we proposed PDD ranking model (2005), and expertise propagation algorithm (2006) which focuses on the social network analysis. In this year, we build different PDD collection from different subsets of the corpus. So each PDD generated has different confidence and reliability. How to estimate the confidence of the expertise and how to merge the ranking lists is what we focus on.

References

1. L. Page, S. Brin, R. Motwani, and T. Winograd (1998). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford, CA.
2. J. M. Kleinberg (1998). Authoritative sources in a hyperlinked environment. Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms.
3. A.P.Dempster, N.M. Laird, and D.B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. J. Royal Statist. Soc. Ser. B., 39, 1977.
4. Y. Fu et al "THUIR at TREC 2005: Enterprise Track" State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University, Proceedings of TREC2005, NIST, 2005
5. D. Ding and B. Zhang. Probabilistic Model Supported Rank Aggregation for the Semantic Concept Detection in Video. ACM International Conference on Image and Video Retrieval (CIVR 2007), July 9-11 2007