The Open University at TREC 2007 Enterprise Track

Jianhan Zhu, Dawei Song, Stefan Rüger
Knowledge Media Institute and Centre for Research in Computing, The Open University, United Kingdom.
{j.zhu, d.song, s.rueger} @open.ac.uk

ABSTRACT

The Multimedia and Information Systems group at the Knowledge Media Institute of the Open University participated in the Expert Search and Document Search tasks of the Enterprise Track in TREC 2007. In both the document and expert search tasks, we have studied the effect of anchor texts in addition to document contents, document authority, url length, query expansion, and relevance feedback in improving search effectiveness. In the expert search task, we have continued using a two-stage language model consisting of a document relevance and co-occurrence models. The document relevance model is equivalent to our approach in the document search task. We have used our innovative multiple-window-based co-occurrence approach. The assumption is that there are multiple levels of associations between an expert and his/her expertise. Our experimental results show that the introduction of additional features in addition to document contents has improved the retrieval effectiveness.

1. INTRODUCTION

In this year's enterprise track, the domain is the website of the CSIRO (Australian Commonwealth Scientific and Research Organization). The task is to find a number of key pages on a topic and a few key experts on the topic in order for a science communicator to create an overview page for the topic. For example, find key experts and key pages on "genetic modification".

Unlike last year's expert search task on the W3C (World Wide Web Consortium) website dataset, expert search on the CSIRO dataset aims to find only a few key contacts on a topic, while expert search on the W3C dataset can find a larger number of experts. Expert search on the CSIRO dataset is judged based on the ground truth provided by science communicators, while expert search results on the W3C dataset were pooled and manually judged by participating groups. Therefore, expert search task this year tends to be more challenging than last year, since the retrieval system needs to not only identify experts on a topic but also rank key contacts among these experts higher than the other non-key contacts.

Another challenge in expert search is that there is not a given list of candidates like in previous two years. This is more like a real world expert search scenario, where there is not a centralized database for maintaining all employees working at an organization. The named entity recognition task gets easier given that all CSIRO staff's email addresses follow the pattern "firstname.lastname@csiro.au". However, one person may have several emails. A mechanism for grouping different emails and name variants under a same person needs to be studied.

A new task in this year's enterprise track is key document search. The task is to identify a few key pages on a topic that a science communicator can put on an overview page about the topic. A key page needs to be not only relevant but also highly authoritative on a topic. This task is similar to the topic distillation task in TREC Web Track. The challenge in document search is how to identify key pages from a large number of documents which are all relevant to the topic on different degrees.

Based on our success in last year's expert search task, we will further investigate the effect of integrating multiple document features in this year's expert and document search.

In both the document and expert search tasks, we have studied the effect of anchor texts in addition to document contents, document authority, url length, query expansion, and relevance feedback in improving search effectiveness and the weighting of the above components in the final document relevance to a topic.

In the expert search task, we have continued using a two-stage model consisting of a document relevance model and a co-occurrence model. The document relevance model is equivalent to our approach in the document search task. We have used our innovative multiple window based co-occurrence approach [3]. The assumption is that there are multiple levels of associations between an expert and his/her expertise. We give higher weights to co-occurrences in smaller windows and lower weights to co-occurrences in larger windows. We have studied different weighting scheme in the multiple-window approach.

In expert name recognition, we have use a clustering algorithm to group email addresses that belong to the same person. We have developed an automatic method for generating variants of an experts' name.

The rest of the paper is organized as follows. We present our document search approach integrating multiple document features in Section 2. A two stage approach consisting of a document relevance model and a co-occurrence model is presented in Section 3. We report our experimental results in Section 4, and conclude in Section 5.

2. DOCUMENT SEARCH

Anchor texts in addition to document contents, document authority, url length, query expansion, and relevance feedback are considered in document search.

2.1 Anchor texts and document content

Anchor texts describe how the others think about a document in a pithy way. We have studied whether anchor texts will improve retrieval results and use different weightings of the contribution of anchor texts and document content in document relevance respectively.

All anchor texts of a document are aggregated together to form an overall anchor text field of the document. A document's relevance to a query is a weighted sum of the relevance of the document's overall anchor text field to the query and the relevance of the document's content to the query. We give higher weight to the anchor text based relevance.

2.2 Inlinks and Outlinks

We study the effect of inlinks and outlinks in document retrieval. Typically, the number of inlinks of a document is an indicator of the document's authority. Previous work shows that there is a strong correlation between the number of inlinks and PageRank [1]. We have combined the number of in-links of a document with the document's content-based relevance. Based on previous work of integrating PageRank in document relevance [2], we have taken the logarithm of the number of inlinks in the combination.

As overview pages on a topic are good candidates and they typically have a relatively large number of outlinks, we have studied whether taking into account outlinks can help improve retrieval effectiveness. Our initial results show that outlinks are not very helpful.

2.3 URL length

The length of the URL of a document shows the depth of the document in the URL hierarchy of a website. Our observation is that authoritative and overview pages on a topic tend to be higher up in the hierarchy. This can be due to various reasons such as that shorter URLs are easy to remember and that document authors tend to assign shorter URLs to key pages which link to a number of pages covering more detailed information on the topic.

We have combined the URL length of a document with the document content-based relevance.

2.4 Query expansion

Narrative part of a topic has been used to enhance document search.

In our automatic runs, a document's relevance to a topic is a weighted sum of the document's relevance to the query part of the topic and the document relevance to the narrative part of the topic.

In our manual runs, the narrative part of a topic was manually modified. A document's relevance to a topic is a weighted sum of the document's relevance to the query part of the topic and the document relevance to the modified narrative part of the topic.

2.5 Relevance feedback

Relevance feedback in terms of using the given key pages to improve the retrieval effectiveness is considered.

3. EXPERT SEARCH

We continue to adopt a two-stage approach in expert retrieval. The two-stage model consists of a document relevance model and a co-occurrence model. The document relevance model is equivalent to the model used in the document search task.

Since the document relevance model has taken into account anchor texts, document authority, url length, query expansion, and relevance feedback, we hypothesize that people appearing in more relevance document are more likely than the other people who do not.

We have continued to use our innovative multiple window based co-occurrence model. A number of windows of different sizes are applied in the co-occurrence model consecutively. The assumption is that there are multiple levels of associations between an expert and a topic, e.g., sentence, paragraph, sec-

tion, …, up to a whole document level. Given a text window, if a person and query terms co-occur, the probability that the person and the topic are associated is higher when the window size is small than the case when the window size is large.

4. NAMED ENTITY RECOGNITION

In expert name recognition from the documents, we use a pattern to find all email addresses ending with ".csiro.au". We will get a large number of email addresses to which we apply a clustering algorithm for grouping email addresses belong to the same person together. The clustering algorithm is based on a similarity measure between each pair of email addresses. The similarity measure is defined based on whether two email addresses share the same last name, the same initials, the same last and first name but one have the middle name but the other does not have the middle name etc.

For each expert, we generate his/her first, last, and possibly middle names based on his/her email addresses. Given the person's name, we generate variants of his/her names. All identifies of a person is matched against the whole corpus for finding out occurrences of the person in the whole corpus.

5. EXPERIMENTAL RESULTS

We have applied our approach to the CSIRO dataset to get four document search runs and four expert search runs for submission. Based our training on the W3C dataset, we have used give incremental text windows for all four runs, i.e., size 5, 20, 80, 200, and 400. Anchor texts, inlinks, and URL length are all considered in the four runs. Descriptions of the four submitted document search runs in Table 1 are as follows.

ouTopicOnly: Only query part of each topic is used in this automatic run.

ouNarrAuto: Narrative part of each topic is used directly in this automatic run. Document relevance to the query part and document relevance to the narrative part are combined for the overall relevance score.

ouNarr: Narrative part of each topic is manually modified in this manual run. Document relevance to the query part and document relevance to the modified narrative part are combined for the overall relevance score.

ouNarrRF: Narrative part of each topic is manually modified in this manual run. Document relevance to the query part and document relevance to the modified narrative part are combined in addition to relevance feedback for the overall relevance score.

Table 1. Document Search Results (The best results for each measure is in bold and underlined)

Runs	MAP	R-Prec	Bpref	P@10
ouTopicOnly	0.3326	0.3734	0.3503	0.5333
ouNarrAuto	0.3137	0.3391	0.3416	0.5238
ouNarr	0.3591	0.3962	0.3682	0.5643
ouNarrRF	**0.3703**	**0.4017**	**0.3793**	**0.5762**

Descriptions of the four submitted expert search runs in Table 2 are as follows.

ouExTitle: Only query part of each topic is used in this automatic run.

ouExNarrAu: Narrative part of each topic is used directly in this automatic run. Document relevance to the query part and document relevance to the narrative part are combined for the overall relevance score.

ouExNarr: Narrative part of each topic is manually modified in this manual run. Document relevance to the query part and document relevance to the modified narrative part are combined for the overall relevance score.

ouExNarrRF: Narrative part of each topic is manually modified in this manual run. Document relevance to the query part and document relevance to the modified narrative part are combined in addition to relevance feedback for the overall relevance score.

Table 2. Expert Search Results (The best results for each measure is in bold and underlined)

Runs	MAP	R-Prec	Bpref	P@10
ouExTitle	0.4337	0.3704	0.8224	0.1560
ouExNarrAu	0.4164	0.3514	0.7851	0.1560
ouExNarr	0.4675	0.4104	0.8391	**0.1640**
ouExNarrRF	**0.4787**	**0.4147**	**0.8457**	**0.1640**

From both Table 1 and 2, we can see that the direct introduction of narrative part in retrieval has negative effective showing that direct use of narrative will introduce more noise than informative keywords. Modified narrative part has help improve both search tasks showing that narrative part contains additional useful information for determining key pages on a topic.

After TREC 2007, we found that there is a mistake in our expert name extraction component for recognizing anti-spam enabled email addresses which results in 73 experts not been recognized in the corpus. After correcting the mistake, we re-ran our experiments with the exact same settings as our four submitted runs, respectively, and got the results shown in Table 3. We can see that the performance of all four runs is largely improved.

Table 3. Re-run expert search results after TREC 2007 (The best results for each measure is in bold and underlined)

Runs	**MAP**	**R-Prec**	**Bpref**	**P@10**
ouExTitle-Re	0.4807	0.3914	0.9118	0.1642
ouExNarrAu-Re	0.4657	0.3742	0.4756	0.1643
ouExNarr-Re	0.5256	0.4480	0.9384	0.1798
ouExNarrRF-Re	**0.5331**	**0.4580**	**0.9391**	**0.1780**

6. CONCLUSIONS

We have participated in both document search and expert search tasks of TREC 2007 Enterprise Track. Our two stage modeling approach has integrated multiple document features in addition to our innovative multiple window based co-occurrence model for effective document and expert search.

ACKNOWLEDGEMENTS

The work reported in this paper is funded in part by an IBM UIMA innovation award and the JISC (Joint Information Systems Committee) funded DYNIQX (Metadata-based DYNamIc Query Interface for Cross(X)-searching content resources) project.

REFERENCES

[1] Chris H. Q. Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, Horst D. Simon (2003) PageRank: HITS and a Unified Framework for Link Analysis. In Proc. of Third SIAM International Conference on Data Mining (SDM).

[2] Ruihua Song, Ji-Rong Wen, Shuming Shi, Guomao Xin, Tie-Yan Liu, et al. (2004) Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004. In Proc. of Text REtrieval Conference (TREC) 2004.

[3] Zhu, J., Song, D., Rüger, S., Eisenstadt, M. and Motta, E. (2006) The Open University at TREC 2006 Enterprise Track Expert Search Task. In Proc. of The Fifteenth Text REtrieval Conference (TREC 2006).