# The Open University at TREC 2006 Enterprise Track Expert Search Task

**Jianhan Zhu, Dawei Song, Stefan Rüger, Marc Eisenstadt, Enrico Motta**
Knowledge Media Institute and Centre for Research in Computing, The Open University, United Kingdom.
{j.zhu, d.song, s.rueger, m.eisenstadt, e.motta} @open.ac.uk

## ABSTRACT

The Multimedia and Information Systems group at the Knowledge Media Institute of the Open University participated in the Expert Search task of the Enterprise Track in TREC 2006. We have proposed to address three main innovative points in a two-stage language model, which consists of a document relevance model and a co-occurrence model, in order to improve the performance of expert search. The three innovative points are based on characteristics of documents. First, document authority in terms of their PageRanks is considered in the document relevance model. Second, document internal structure is taken into account in the co-occurrence model. Third, we consider multiple levels of associations between experts and query terms in the co-occurrence model. Our experiments on the TREC2006 Expert Search task show that addressing the above three points has led to improved effectiveness of expert search on the W3C dataset.

## 1. INTRODUCTION

The aim of this year's expert search is to find W3C people in the expert candidate list who are the best matches for a user's specific expertise request on a topic. For example, a user is looking for an expert "who has knowledge in relationship cardinalities between roles in different choreographies". The expertise request is often complex and needs to be converted into queries for IR systems to process. Taking a starting query as the input, query expansion and formulation methods can generate more precise and informative queries for better retrieval results.

In last year expert search task, several systems have followed a language modeling approach and have consistently achieved good results [3, 4]. Thus we have also adopted the two-stage language model approach. The two-stage language model consists of a document relevance model and a co-occurrence model. First, the document relevance model finds documents which are relevant to the expertise topic. Second, a co-occurrence model is used to find experts who are closely related to the expertise topic based on the assumption that if an expert's identity (such as his/her name, email address, user id) co-occurs with the terms of a query describing the topic in a text window, the expert is likely to be related to the topic.

We have proposed three innovative points to improve the two-stage language model approach.

First, Google has used PageRanks of documents to combine with content-based document relevance in finding authoritative documents on a query, and has greatly improved the user's search experience on the web. We think that document authority should be taken into account in the document relevance model as well. By doing so, experts from these authorities can be correctly identified.

Second, many documents on an organizational intranet are semi-structured. For example, a W3C technical report can often be segmented into the title, author, editor, co-authors, various sections, references, acknowledgements, and appendixes parts. The occurrences of an expert in different parts of a document will affect the co-occurrence model. We propose to take into account the document internal structure in the co-occurrence model.

Third, in typical window-based association methods, a text window is set to measure the co-occurrences of an expert and query terms. Once the window size is set, it is fixed. However, in expert search, there are associations between an expert and query terms on multiple levels, i.e., from phrase, sentence, paragraph, etc., up to document levels. All these levels of associations need to be considered in the co-occurrence model. Increased window sizes often lead to more coverage of associations while introducing noise. We propose to adopt a weighted multiple window size approach in the co-occurrence model.

We propose to integrate the above three innovative points in a two-stage language model for more effective expert search than using document content alone. To the best of our knowledge, we are the first to use a weighted-multiple-window-based approach in a language model for association discovery. In the rest of the paper, we will discuss the query expansion and formulation in Section 2. Expert identity recognition is covered in Section 3. The three inno-

vative points, i.e., document authority in the document relevance model, document internal structure in the co-occurrence model, and weighted-multiple-window-based approach in the co-occurrence model, are discussed in Section 4, 5, and 6, respectively. The overall two-stage language model is presented in Section 7. Our experimental results on the W3C dataset are reported in Section 8. Finally, Section 9 concludes the paper.

## 2. QUERY EXPANSION AND FORMULATION

A searcher's expertise request on a topic can often be complex and needs to be converted to queries for IR systems to process. For the example in Paragraph 1 of Section1, a narrative of the topic is "in the context of semantic web, the relationships between entities can have different cardinalities and roles. Relevant expert will have an explicit knowledge of such choreographies. Experts in Semantic web are not relevant without explicit knowledge in choreographies."

The title of a topic is typically used as the search query. However, by considering the description and narrative parts of the topic, this query can often be expanded to form a number of more precise and informative queries leading to better search results. For the example in the pervious paragraph, the title "relationship cardinalities" can be expanded for a number of queries such as "relationship cardinalities choreography" and "relationship cardinalities role" etc.

There are two ways in query expansion and formulation, i.e., automatic methods and manual methods. Automatic methods such as the Robertson Term Selection in BM25 [6] and HAL-based information inference methods [8] are used for query expansion and term weighting in our experiments. Automatically generated queries are manually viewed, selected and modified by a human expert.

## 3. EXPERT IDENTITY RECOGNITION

A list of experts' names and their email addresses is provided for identity recognition in the task. However, variants of experts' names need to be considered to improve the recall of the recognition. An automatic method can be used to generate the typical variations of a person's name, e.g., given "Deborah L. McGuinness", the automatically generated variants are "Deborah McGuinness", "McGuinness, Deborah L.", "McGuinness, D. L." etc. Typical correspondence between real names and nicknames are made, e.g., "Michael" and "Mike", "Deborah" and "Deb". Special attention is paid to the European names with non-English letters. Conventional correspondence is made between European letters and English letters, e.g., ë →e, ø→oe. We pre-process the documents by removing HTML mark-up etc., and use the Aho-Corasick algorithm [2] to match these expert identities against the pre-processed documents.

## 4. DOCUMENT AUTHORITY IN DOCUMENT RELEVANCE MODEL

In an organizational intranet such as the W3C website, some documents are more authoritative than the others in identifying people's expertise on a topic, thus giving higher weights to these authorities than the other ordinary documents can potentially improve expert search performance. These authorities typically are linked to many other external authoritative sources. We have experimented with using Google rankings of documents, which incorporate PageRank measuring page authority, to identify authoritative documents and give them higher weights in expert search. We also gave higher weights to authoritative mailing lists on a topic identified by their URL prefixes and judged by their Google rankings.

## 5. DOCUMENT INTERNAL STRUCTURE IN CO-OCCURRENCE MODEL

A document's internal structure can often be crucial in determining whether a person mentioned in the document is an expert on a topic also mentioned in the document. For example, in a W3C technical report, the occurrence of a person's name in the editor, author, content, reference, or acknowledgement section of the report has different implications of the person's expertise on a topic. Since documents on the W3C site often follow a certain pattern in formatting their structures, such pattern can be discovered and used to segment these documents into multiple sections. We have experimented with taking into account the internal structures of technical reports, emails, and papers, e.g., if a person appears in the author section of a report, we will give high weight to the person's relevance to a topic in the report; if he/she appears in the reference section, his/her relevance to a topic co-occurring in a small window is given high weight, but is given low weight when they co-occur in a large window.

## 6. INCREMENTAL WINDOW SIZES IN CO-OCCURRENCE MODEL

In selecting window sizes, small window sizes often lead to high precision but low recall in finding experts relevant to a topic, while large window sizes lead to high recall but low precision. We used incremental window sizes and took a weighted sum of their respective expert-to-topic association scores to measure an expert's overall association to a topic in a document. We gave high weights to association based on small windows, and low weights to association based on large windows. The window sizes were chosen to reflect from phrase level, sentence level, paragraph level, section level, etc., up to document level relevance.

## 7. TWO-STAGE LANGUAGE MODEL

The two-stage model consists of a document relevance model and a co-occurrence model. Given the topic, an expert (*e*)'s relevance to the topic *t* is denoted *P(e|t)*. A number of queries, each of which is denoted as $q_i$, are expanded and formulated from *t* for expert search, and $P(e|t)=\sum_i P(q_i)\cdot P(e\,|\,q_i)$, where $P(q_i)$ is the prior probability of $q_i$ indicating how much we want to see query terms in $q_i$ co-occur with *e* in relative to other queries. The two-stage language model is as follows:

$$P(e|q_i)=\sum_d P(e,d\,|\,q_i)=\sum_d P(d\,|\,q_i)P(e\,|\,d,q_i)$$

Where $P(d|q_i)$ is the document relevance model, and $P(e|d,q_i)$ is the co-occurrence model.

We have experimented with different methods in the document relevance model. Span query, Boolean query, and BM25 are used. Span query can be seen as a co-occurrence model where all query terms are required to co-occur in a text window. Boolean query specifies that all query terms must occur in a document. BM25 is a probabilistic model [7].

PageRanks of documents are later combined with the document relevance. Craswell et al. [5] proposed a method for combining PageRanks with BM25. Due to the difficulty of calculating PageRanks of the W3C documents, an alternative approach of using Google rankings of documents as an estimate of their PageRanks is adopted. The equation is:

$$P(d|q_i)= P_{content}(d|q_i) +w*1/GoogleRank(d|\,q_i)$$

Where *w* is the weight given to the document authority part.

In the co-occurrence model, both document internal structure and incremental window sizes are considered. We have used Lucene's [1] span query to achieve this. The equation is:

$$P(e\,|\,d,q_i)=\sum_{k=1,\ldots,M}\sum_W F(W,Section(e_k))\cdot P_W(e_k\,|\,d,q_i)$$

Where $F(W, Section(e_k))$, a function of the window size and the section where $e_k$ occurs, is the weight for the association score between the *k*th occurrence of expert *e* and the query in *d*. $P_W(e_k\,|\,d,q_i)$ is given by the span query score in Lucene which takes into account span frequency and document span frequency similar to the *TF/IDF* measure.

## 8. EXPERIMENTAL RESULTS

We have applied our approach to a subset of the W3C dataset to get four submitted runs. The subset consists of the WWW, Lists, People, ESW, and Other sections. An initial text window size is set as 10 for all four runs. 10 incremental text windows are used for all four runs, i.e., size 10, 28, 48, 88, 160, 280, 360, 660, 1200, and 3200. Query expansion, internal structure, and document authority are considered in all four runs. Descriptions of the four submitted runs are as follows:

**kmiZhu1**: Boolean query is used for document relevance. Document relevance is aggregated on the basis of the number of spans.

**kmiZhu2**: Span query is used for document relevance. Document relevance is aggregated on the basis of the number of spans.

**kmiZhu4**: Span query is used for document relevance. Document relevance scores are aggregated on a per document basis instead of the number of spans.

**kmiZhu5**: Boolean query is used for document relevance. Document relevance is aggregated on the basis of the number of spans. Use lower weights than the other three runs for parts of a document that are identified as more important than the other parts, e.g., the author section of a technical report, in the co-occurrence model.

The four runs' retrieval results without and with taking into account supporting documents are shown in Table 1 and 2, respectively.

**Table 1**. Retrieval performance without considering supporting documents.

| Runs | MAP | R-precision | Bpref | P@10 |
|---|---|---|---|---|
| kmiZhu1 | **0.6431** | 0.6242 | **0.6391** | 0.7347 |
| kmiZhu2 | 0.6329 | 0.6078 | 0.6265 | 0.7408 |
| kmiZhu4 | 0.6385 | 0.6121 | 0.6237 | **0.7551** |
| kmiZhu5 | 0.6401 | **0.6267** | 0.6369 | 0.7367 |

**Table 2**. Retrieval performance considering supporting documents.

| Runs | MAP | R-precision | Bpref | P@10 |
|---|---|---|---|---|
| kmiZhu1 | **0.4421** | 0.4835 | **0.4986** | 0.5633 |
| kmiZhu2 | 0.4402 | 0.4650 | 0.4911 | 0.5857 |
| kmiZhu4 | 0.4155 | 0.4503 | 0.4700 | **0.5694** |
| kmiZhu5 | 0.4416 | **0.4852** | 0.4983 | 0.5673 |

From Table 1, we can see that the run kmiZhu1 achieved the highest MAP and Bpref, kmiZhu4 achieved the highest P@10, and kmiZhu5 got the best score in R-precision. The following points can be learned from Table 1:

1. We can see that using span query does not lead to better performance compared with the Boolean query.
2. Aggregating span query based document relevance scores on a document basis leads to better results than that on a span basis, revealing that the span query based document relevance has already taken the number of spans in a document into account and the number of spans does not need to be considered twice.
3. Run kmiZhu4 achieved the best P@10. We can see that span query based method has the advantage of identifying more correct experts in the top 10 results than the Boolean query method.
4. Run kmiZhu5 with lower weights for important parts of documents in the co-occurrence model has worse performance than run kmiZhu1. However, we will carry out experiments to find the best weighting scheme, which can maximize the MAP of retrieval results.

Since documents supporting a person's expertise on a topic are also retrieved, a person can only be judged as an expert when he/she is judged as an expert on his/her own right with at least one supporting document also retrieved. The results are shown in Table 2.

We can see that kmiZhu1 is still the best run in terms of MAP and Bpref, run kmiZhu4 is the best in P@10, and kmiZhu5 got the best score in R-precision. However, we can see that there is a considerable drop of performance in terms of all runs compared with the results in Table 1. The need to improve the current method of simply using the document relevance model to select the top 20 supporting documents is necessary.

We have also done experiments to find out the contribution of each component in our approach in expert search, and our findings are as follows:

1. We used BM25 as the document relevance model to substitute the document relevance model in the four submitted runs, and got slighter lower MAPs than the four submitted runs, respectively.
2. We used a baseline, which uses a single fixed window without considering query expansion, internal structure, and document authority. We then added query expansion, internal structure, document authority, and multiple windows to the baseline, respectively. We found that query expansion helped the performance of the baseline increase greatly. Internal structure can contribute to increase of the baseline in terms of MAP. We have used the Google based document rankings, and two set of PageRanks contributed by Danil Nemirovsky and SJTU, respectively, hosted by NIST (http://ir.nist.gov/w3c/contrib/), and found that document authority made little or no contribution to the baseline in terms of MAP. Using multiple windows can help increase the MAP of the baseline, but the selection of these windows needs further investigation.

## 9. CONCLUSIONS AND FUTURE WORK

In TREC2006 Enterprise Track Expert Search task, our best run has achieved the MAP of 0.6431 and 0.4421 for without and with considering supporting documents, respectively. Our best run is also the most effective among all the runs submitted by the 23 participating groups in the task in terms of MAP, R-precision, Bpref, and P@10. Our innovative approach of integrating three document characteristics, namely, multiple levels of associations between experts and query terms, document internal structure, and document authority, in a two-stage language model for expert search has greatly improve the per-

formance of a baseline two-stage language model which uses the document content alone.

Based on the test collection built this year, future work includes:

1. A systematic investigation of the influence of document authority, document internal structure, and multiple window sizes on the expert search retrieval results, e.g., trying out different methods for incorporating document authority in document relevance, experimenting with different weight methods for document internal structures and multiple window sizes, and investigating efficient methods for window combination optimization etc.
2. Improve the method for identifying expertise supporting documents.

**ACKNOWLEDGEMENTS**

The work reported in this paper is funded in part by the Advanced Knowledge Technologies (AKT) project[1] and an IBM UIMA innovation award.

**REFERENCES**

[1] http://lucene.apache.org

[2] Aho, A. V., Corasick, M.J. (1975) Efficient string matching: An aid to bibliographic search. Communications of the ACM 18 (6): 333–340.

[3] Balog, K., Azzopardi, L. and de Rijke, M. (2006) Formal models for expert finding in enterprise corpora. In Proc. of SIGIR, 43-50.

[4] Cao, Y., Liu, J., Bao, S. and Li, H. (2005) Research on Expert Search at Enterprise Track of TREC 2005. In Proc. of TREC 2005.

[5] Craswell, N., Robertson, S.E., Zaragoza, H., and Taylor, M.J. (2005) Relevance weighting for query independent evidence. SIGIR 2005: 416-423.

[6] Robertson, S.E. (1990) On term selection for query expansion. Journal of Documentation 46, 359-364.

[7] Robertson, S.E. and Jones, K.S. (1994) Simple, proven approaches to text retrieval. University of Cambridge Computer Laboratory Technical Report no. 356.

[8] Song, D. and Bruza, P.D. (2003) Towards Context Sensitive Informational Inference. Journal of the American Society for Information Science and Technology (JASIST), 52(4), 321-334.

[1] AKT is an Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University.