The University College London at TREC 2008 Enterprise Track

Jianhan Zhu
University College London, Adastral Park Campus, Ipswich, IP5 3RE, U.K.
Email: j.zhu@adastral.ucl.ac.uk

ABSTRACT

The University College London Information Retrieval Group participated in both the Expert Search and Document Search tasks in the TREC2008 Enterprise Track. We used a generic two-stage approach, which consists of a document retrieval stage followed by an expert association discovery stage, for expert finding. Since document search is an integral part of our expert finding approach, we have studied the relationship between document search and expert search. Due to the existence of rich features that can potentially contribute to expert finding, our expert finding approach integrates these features including anchor texts, indegree, and multiple levels of associations between experts and query terms. Our experimental results show that the introduction of features has helped improve the expert finding performance.

1. INTRODUCTION

Same as in TREC2007 Enterprise Track, the domain for TREC2008 Enterprise Track is the website of the CSIRO (Australian Commonwealth Scientific and Research Organization). The topics were developed in order to reflect the requests of information received by the CSIRO Enquiries staffers. The aim of the two tasks is to find a number of key pages and experts on a topic that can help the staffers to answer each request. For example, find key experts and key pages to answer the request for information on "cane toad".

Based on our approach that integrates multiple features in a two-stage expert finding model [4], we have continued investigating the effects of these features as follows in expert finding.

Anchor texts: anchor texts of a document often highlight its key topic. Sometimes, keywords for identifying a document's topic may even be missing in the document itself but exist in its anchor texts, e.g. the BMW homepage does not mention "car", but anchor texts pointing to the page often do. We have studied the effect of anchor texts in both expert and document search.

Indegree: 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], and PageRank and indegree help document search on the Web [2]. We will study the effect of indegree in both document and expert search.

Multiple levels of associations: We have continued using our multiple window based co-occurrence model [4]. The assumption is that there are multiple levels of associations between an expert and query terms in documents. We give higher weights to co-occurrences in smaller windows and lower weights to co-occurrences in larger windows. We have studied different window selections and combinations.

In [3], we studied the relationship between ad hoc retrieval and expert finding via three parameters, namely, a background smoothing parameter in a language model, and anchor texts and indegree. Our experiments on the TREC 2007 Enterprise Track CSIRO dataset have shown that improvement in document retrieval does not necessarily lead to improvement in expert finding.

Firstly, smoothing language model by a background collection model can significantly improve ad hoc retrieval performance, but does not help or even hurt expert finding. Accordingly, we give background smoothing different weights for expert and document search, respectively.

Secondly, anchor text does not help document retrieval, and hurts document retrieval when weighted high in document retrieval, and indegree only slight helps ad hoc retrieval. Therefore, anchor texts and indegree have different effect in intranet search than in Web search [2].

The reason might be that, in document retrieval, documents are largely judged as relevant or not regardless of their authoritativeness, and anchor text and indegree may introduce more noise than useful information in document retrieval.

However, both anchor text and indegree help expert finding. Since people appearing in authorita-

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The University College London at TREC 2008 Enterprise Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University College London,Adastral Park Campus,Ipswich, IP5 3RE, U.K.,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

tive documents are more likely to be experts than those appearing in ordinary documents, anchor text and indegree, which bias towards authoritative documents, can help expert finding. Therefore, we used anchor text and indegree for expert search, but not for document search.

The rest of the paper is organized as follows. We present our two-stage approach for expert fining in Section 2, report experimental results in Section 3, and conclude in Section 4.

2. TWO-STAGE EXPERT FINDING

Given a set of documents *d*, a query topic *q*, and a set of candidates *c*, the aim of expert finding is to estimate $p(c|q)$ for ranking the candidates. Since $p(c|q) = p(c,q)/p(q)$ and $p(q)$ does not affect ranking, the task is to estimate $p(c,q)$.

We adopt a document-centric generative approach, and represent the joint as a weighted average of the document models as:

$$p(c,q) = \sum_d p(c,q \mid d)p(d) = \sum_d p(c \mid q,d)p(q \mid d)p(d) \quad (1)$$

The document prior $p(d)$ is estimated by the indegree of *d*, and $p(d) \propto f_{indegree}(d)$, where $f_{indegree}(d)$ is the transformation function for indegree.

We use Craswell et al. [2]'s *sigm* transformation function for estimating $f_{indegree}(d)$:

$$f_{indegree}(d) \propto w\frac{indegree(d)^a}{k^a + indegree(d)^a} \quad (2)$$

where *w*, *a* and *k* are parameters, and *indegree(d)* is the indegree of *d*. We use the same parameters that were used in [2], and set the values of *w*, *a* and *k* as 3.7, 0.2, and 5 respectively.

$p(q|d)$ is estimated by inferring a document language model θ_d for each document *d* as

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)^{n(t,q)} \quad (3)$$

where *t* is a query term and $n(t,q)$ is the number of times it is used in *q*. We smooth the document language model with the background model, and take into account anchor texts by using a mixture of document content and anchor text to represent each document, therefore

$$p(t \mid \theta_d) = (1-\lambda_c)(\lambda_t p(t \mid d_{text}) + \lambda_a p(t \mid d_{anchor})) + \lambda_c p(t) \quad (4)$$

where the document content part is weighted with $(1-\lambda_c)\lambda_t$, anchor text part is weighted with $(1-\lambda_c)\lambda_a$, $\lambda_t+\lambda_a = 1.0$, and $p(t)$ is the maximum likelihood estimate of the term *t* given the background model.

We used Dirichlet smoothing for adjusting λ_c.

$$\lambda_c = \frac{\mu}{|d| + \mu} \quad (5)$$

where |d| is the length of document *d*.

In Eq. 1, $p(c|q,d)$ denotes a co-occurrence model which is constructed as a linear interpolation of $p(c|q,d)$ and the background model $p(c)$ to ensure there are no zero probabilities, we get

$$p(c \mid \theta_d, \theta_q) = (1-\mu_c)p(c \mid q,d) + \mu_c p(c) \quad (6)$$

where $p(c)$ is the probability of candidate *c*. We estimate $p(c)$ as

$$p(c) = \frac{1}{df_c}\sum_{d'} \frac{f(c,d')}{\sum_{c' \in C} f(c',d')} \quad (7)$$

where $f(c,d')$ is the frequency of candidate *c* in document *d'* and df_c is the document frequency of *c*.

We use a Dirichlet prior for the smoothing parameter μ_c

$$\mu_c = \frac{\kappa}{\sum_{c'} f(c',d') + \kappa} \quad (8)$$

where κ is the average term frequency of all candidates in the corpus.

We use a multiple window based approach in estimating $p(c|d,q)$. We assume that small windows often lead to more probable associations, and large windows result in nosier associations, and weight smaller windows higher than larger ones.

Given a list *W* consisting of *N* windows $\{w\}$ of different sizes, we estimate $p(c|d,q)$ as

$$p(c \mid d,q) = \sum_w p(w)p(c \mid d,q,w) \quad (9)$$

where $p(w)$ is the probability for each of the window-based co-occurrence models.

Given a number of text windows of size *w* where *c* co-occurs with *q* as $\{w_i\}$, we estimate $p(c|q,d,w)$ as follows

$$p(c \mid d,q,w) = \sum_{w_i} \frac{f(c,d,q,w_i)}{\sum_{c'} f(c',d,q,w_i)} \quad (10)$$

where $f(c,d,q,w_i)$ is the frequency of c in a text window, and $\sum_{c'} f(c',d,q,w_i)$ is the total frequency of candidates in the window.

3. EXPERIMENTAL RESULTS

There is not a given list of candidates. By utilizing the pattern that most of the CSIRO staff's email addresses follow the pattern

"firstname.lastname@csiro.au", we have extracted a list of candidates. By considering that one person may have several emails and aliases, we developed a method for aligning different emails and name variants.

Document search and expert search are two important tasks in an enterprise environment. Based on our previous findings on the TREC2007 CSIRO collection, we adapt our approach to document and expert search, respectively.

Firstly, based on the finding that background smoothing is very helpful to document search but may harm expert finding when the background smoothing is too much, we used the Dirichlet smoothing language model where the smoothing parameter μ is given a small value, such as 100, for expert search.

Secondly, based on the finding that anchor texts and indegree are more helpful to expert search than document search, we did not incorporate anchor texts and indegree in document search, but instead integrated them in our expert finding approach.

Based on the above decisions, we submitted four document search runs, and four expert search runs, and the results are summarized in Table 1 and 2.

Descriptions of the four submitted document search runs are as follows.

Ucl01: Title only automatic run, we used Dirichlet smoothing language model where the smoothing parameter μ is set as 2000.

Ucl02: Title only automatic run, we used Dirichlet smoothing language model where the smoothing parameter μ is set as 3000.

Ucl03: Title only automatic run, we used Dirichlet smoothing language model where the smoothing parameter μ is set as 2500.

Ucl04: Title only automatic run, we used the BM25 model where the parameters K_1 is 1.4, b is 0.6, and K_3 is 8.

We varied the Dirichlet smoothing parameter, and compared the Dirichlet smoothing model with the BM25 model for document search. We can see from Table 1 that the Dirichlet smoothing where the parameter is set as 2000 leads to the best performance on both infAP and infNDCG metrics. The BM25 model performed slightly worse than the Dirichlet smoothing language model

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

Runs	Ucl01	Ucl02	Ucl03	Ucl04
infAP	**0.3246**	0.3158	0.3205	0.3031
infNDCG	**0.5175**	0.5141	0.5172	0.4965

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

UCLex01: Window size 450, anchor text, and indegree

UCLex02: Window size 600, anchor text, and indegree.

UCLex03: Multiple windows 40, 400, and 800, anchor text, and indegree.

UCLex04: Multiple windows 40, 200, 400, and 800, anchor text, and indegree.

Firstly, we have compared the effect of different window sizes, and the run UCLex01 with window size 450 outperforms the run UCLex02 with window size 600 in terms of the MAP, MRR, and P@5. This shows that smaller windows lead to more accurate associations, and larger windows may introduce more noise. However, the run UCLex02 has achieved higher num_rel_ret and P@100 than the run UCLex01, showing that larger windows can discover more expert associations with the queries.

Secondly, we have compared multiple windows with single windows, and different window combinations. Our two multiple window based runs both outperformed the two single window based runs, showing that weighting expert associations based on windows sizes can help improve expert finding performance. A finer-grained multiple window approach, i.e., UCLex04, achieved the highest performance on a number of metrics including MAP, R-Prec, Bpref, P@5, and P@10.

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

Runs	UCLex01	UCLex02	UCLex03	UCLex04
MAP	0.3360	0.3346	0.3433	**0.3476**
MRR	**0.6789**	0.6737	0.6748	0.6759
Num_rel_ret	335	342	**347**	346
R-prec	0.3332	0.3340	0.3330	**0.3378**
Bpref	0.3740	0.3782	0.3781	**0.3816**
P@5	0.4400	0.4327	0.4364	**0.4473**
P@10	**0.3164**	**0.3164**	0.3145	**0.3164**
P@100	0.0609	0.0622	**0.0631**	0.0629

4. CONCLUSIONS

We have participated in both document and expert search tasks of TREC 2008 Enterprise Track. We

have continued using a two-stage expert finding approach which integrates features including anchor texts, indegree, and multiple levels of associations. Our submitted runs to TREC2008 have shown the effectiveness of multiple windows and effect of window selections. Document search is an integral part of our expert finding approach. Based on our previous findings that background smoothing is helpful to document search but may hurt expert finding, and anchor texts and indegree are both helpful to expert finding but less helpful to document search, we adapted our approach to expert and document search, respectively.

REFERENCES

[1] Upstill, T., Craswell, N., and Hawking, D. (2003) Predicting Fame and Fortune: PageRank or Indegree? In Proc. of the Australasian Document Computing Symposium ADCS 2003.

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

[3] Zhu, J. (2008) A study of the relationship between ad hoc retrieval and expert finding in enterprise environment. In Proc. of the CIKM workshop on Web Information and Data Management (WIDM) 2008, pp. 25-30

[4] Zhu, J., Song, D., Rüger, S., Eisenstadt, M. and Motta, E. (2007) The Open University at TREC 2006 Enterprise Track Expert Search Task. In Proc. of the Fifteenth Text REtrieval Conference (TREC 2006).