A study of the relationship between ad hoc retrieval and expert finding in enterprise environment

Jianhan Zhu
University College London
Adastral Park Campus, Ipswich
IP5 3RE, United Kingdom
j.zhu@adastral.ucl.ac.uk

ABSTRACT
Ad hoc retrieval returns a ranked list of documents in response to a search query, while expert finding returns a ranked list of people in response to an expertise request in the form of a search query, e.g., "information retrieval". In current state of the art expert finding approaches, ad hoc retrieval is a key component for locating documents relevant to the expertise request. While ad hoc retrieval has been well researched in information retrieval, no previous work has been carried out on the effects of document retrieval in expert finding. The main contribution of this paper is that we are the first to study the effect of document retrieval in expert finding via a background smoothing parameter in a language modeling approach and two document features, namely, anchor text and indegree. Our research gives insight into how to design an effective approach for both ad hoc retrieval and expert finding in enterprise environment. Our experiments on the TREC (Text REtrieval Conference) 2007 Enterprise Track CSIRO (Australian Commonwealth Scientific and Research Organization) dataset shows that background smoothing helps improve ad hoc retrieval but does not help or even hurt expert finding, anchor text helps expert finding but hurt ad hoc retrieval when weighted high, and indegree helps expert finding but does not help improve ad hoc retrieval significantly.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H3.1 Content analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms
Experimentation, Measurement, Performance

Keywords
Ad hoc retrieval, expert finding, language models

1. INTRODUCTION AND MOTIVATION
Expert finding is a key task in enterprise search and has recently attracted lots of attention from both research and industry communities as evidenced by the organization of expert search tasks in the Text REtrieval Conference (TREC) track [1, 7, 15]. A typical user scenario is one in which users need to learn about a subject and want to talk to someone who knows about it as the first step. Another use case is when a project manager wishes to assemble a project team made of people with a range of skills [19]. Yimam-Seid and Kobsa [17] identified two main motives for expert finding, namely, as a source of information to answer the question "who knows about topic x?" (i.e., to find experts for a particular topic) and also to answer questions such as "does person y know about topic x?" or "what else does y know?"

The TREC enterprise track [1, 7, 15] has been the major forum for empirically comparing expertise modeling techniques. Essentially, the two most popular and well-performing types of approaches in TREC expert search task are profile-centric and document-centric approaches [1, 7, 15].

Expert-profile-centric approaches build an expert profile as a pseudo document by aggregating text segments relevant to the expert, e.g., context text windows of the expert in documents [10]. Profiles of experts as pseudo documents are indexed and ad hoc information retrieval models are applied to the index for finding experts on a topic.

Document-centric approaches are typically based on ad hoc information retrieval techniques. This type of approaches models how users will typically search for experts using an ad hoc IR system, such as a search engine like Google. The user first finds a number of documents relevant to a search topic, such as "information retrieval", via a search engine and then read these documents in order to find a number of experts on the topic based on the context and the frequency of co-occurrences of these people and query terms. For example, people who often appear in the context of "information retrieval" in many relevant documents are likely to be experts.

We can see that ad hoc retrieval plays an important role in document-centric expert finding approaches. On the other hand, both ad hoc retrieval and expert finding are important tasks in an organizational environment. Our work addresses the challenge of how to design effective approaches for both ad hoc retrieval and expert finding.

Typically, good expert finding results rely on the retrieval of documents highly relevant to the search topic. But does more effective ad hoc retrieval always lead to better expert fining results, and vice versa? This is the research question we would like to answer in this paper.

In this paper, we use language models [13] for ad hoc retrieval, since they are the state of the art models for ad hoc retrieval with consistent and good performance on TREC collections [13, 18].

Let us model a generic document-centric expert finding approach as follows:

Firstly, in a language model for ad hoc retrieval, we estimate the conditional probability $p(q|d)$, of a query topic q given a document d. Based on the assumption that terms co-occurring with an expert in the same context describe the expert, $p(q|d)$ is used to weight the evidence of co-occurrences of experts with terms in q in documents. The conditional probability $p(c|q)$ of an expert candidate c given a query q can be estimated by aggregating all the evidences in all the documents where c and terms in q co-occur.

Although ad hoc retrieval is an important first step for document-centric expert finding, no previous work has been carried out on the relationship between ad hoc retrieval and expert finding. In this paper, we will study the effect of ad hoc retrieval in expert finding via a background smoothing parameter in a language model approach, and two document features, i.e., indegree and anchor texts.

Background smoothing parameter: in language model approach, a multinomial distribution $p(w|D)$ for each term w is estimated for each document D in a collection C. This model is used to assign a likelihood to a user query $q=(t_1,t_2,\ldots, t_m)$, where t_i $(1\leq i\leq m)$ are terms. Assuming query terms are independent, $p(q|D)$ is the product of $p(t_i|D)$ [13]. Language model for ad hoc retrieval must be "smoothed" so that non-zero probability can be assigned to query terms that do not appear in document D. Usually, document language model is smoothed by linear interpolation with a background collection model $p(w|C)$. The effect of the background model can be adjusted by a smoothing parameter, which will be discussed in details in Section 3.

Indegree: the number of incoming links of a document (indegree) correlates with the document's PageRank [16]. Craswell et al [6] integrate PageRank and indegree with a BM25 baseline model for more effective document retrieval than the BM25 baseline model. Cheng et al [4] propose the use of PageRank for entity retrieval. We hypothesize that more authoritative documents are typically linked to more often by other documents. Based on the assumption that people mentioned in authoritative documents are more likely to be experts on a topic, we will investigate the effect of indegree in ad hoc retrieval and 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. Anchor texts have been shown to be helpful in document retrieval on the Web [5, 8]. We will study whether the effectiveness of anchor texts in document retrieval can be converted into their effectiveness in expert finding.

Previous work on ad hoc retrieval shows that the collection background model helps improve retrieval effectiveness dramatically [13, 18], and the two document features also help improve document retrieval on the Web [5, 6, 8]. We will study whether more effective ad hoc retrieval leads to better expert finding results, and vice versa, via background smoothing, and two document features.

The rest of the paper is organized as follows. In Section 2, we introduce related work on expert finding. We present our two-stage language modeling approach for expert finding in Section 3. We report our experimental results on the TREC2007 CSIRO dataset for both ad hoc retrieval and expert finding in Section 4. Finally, we conclude in Section 5.

2. RELATED WORK

Previous research on expert finding shows that document-centric approaches normally outperform profile-centric approaches [2, 15] as the latter achieves efficiency at the expense of useful information in terms of internal document structure and high-level language features [11].

Cao et al. [3] first propose a two-stage language model combining a document relevance and co-occurrence model. Fang et al [9] derive a generative probabilistic model from the probabilistic ranking principle and extend it with query expansion and non-uniform candidate priors. A prominent language modeling approach has been proposed by Balog et al [2]. They distinguish between "Model 1", which directly represents the knowledge of an expert from associated documents, and "Model 2", which first locates documents on a topic and then finds the associated experts. Their experiments on TREC dataset show that "Model 2" outperforms "Model 1" in expert finding. Their "Model 1" is equivalent to a profile-centric approach, and "Model 2" equivalent to a document-centric approach. Petkova and Croft [12] have further improved their models by proposing a proximity-based document representation for incorporating sequential information in text. We [20] first propose a novel multiple window based approach for integrating multiple levels of associations between experts and a query in expert finding. Serdyukov and Hiemstra [14] propose a novel expert-centric language model for expert search.

We study the effects of a number of document features in expert finding in [21]. These document features include document internal structure, document URL length, PageRank, indegree, anchor text, and multiple levels of associations between experts and topics. The work reported in this paper will explore the relationship between ad hoc retrieval and expert finding via a number of features.

3. LANGUAGE MODEL FOR AD HOC RETRIEVAL AND EXPERT FINDING

Our models are instances of document-centric generative language modeling approaches to rank experts. Given a set D of documents, a query topic q, and a set C of candidates, the aim is to determine $p(c|q)$ and rank the set of candidates according to this probability: $p(c|q) = p(c,q)/p(q)$, where $p(c,q)$ is the joint probability of the candidate and query, and $p(q)$ is the probability of the query q. Assuming that $p(q)$ is uniform implies that $p(c|q)$ is proportional to $p(c,q)$.

To determine $p(c,q)$, we adopt a document-centric generative language modeling approach. We randomly draw independent samples of documents from $p(c,q)$ and represent the joint as a weighted average of the document models in Equation 1.

$$p(c,q) = \sum_{d \in D} p(c,q \mid d) p(d) = \sum_{d \in D} p(c \mid q,d) p(q \mid d) p(d) \quad (1)$$

We estimate $p(q|d)p(d)$ for ad hoc retrieval. Most previous approaches ignore the document prior $p(d)$ by assuming that it is uniform for all documents. However, we estimate $p(d)$ based on the indegree of d, therefore $p(d) \propto f_{indegree}(d)$, where $f_{indegree}(d)$ is the transformation function for indegree.

Craswell et al [6] use a number of query independent features including PageRank, indegree, URL Length and ClickDistance for ad hoc retrieval. They propose sigmoid transformations of these features in combination with a BM25 baseline. Their experiments on TREC Web Track data show the usefulness of these features in ad hoc retrieval.

We use Craswell et al. [6]'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 [6], 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 [13] 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, i.e., the whole collection, 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 $(1\text{-}\lambda_c)\lambda_t$, anchor text part is weighted with $(1\text{-}\lambda_c)\lambda_a$, $\lambda_t + \lambda_a = 1.0$, and $p(t)$ is the maximum likelihood estimate of the term t given the background model, weighted with λ_c.

We used Dirichlet smoothing which is also called Bayesian smoothing [18] for adjusting λ_c.

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

where $|D|$ is the length of document D.

In Equation 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

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

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

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

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

We estimate $p(c|q,d)$ as follows

$$p(c \mid q,d) = \frac{f(c,d,q)}{\sum_{c' \in C} f(c',d,q)}, \quad (9)$$

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

4. EXPERIMENTAL RESULTS

We carried out our experiments on the TREC2007 CSIRO dataset, which is a crawl of the publicly available web pages from the *.csiro.au domain, known as the CSIRO Enterprise Research Collection [1] (http://es.csiro.au/cerc/). The dataset consists of 370,715 documents with an average document length of 457.01 terms and 1,549,127 unique terms altogether.

Based on the observation that most CSIRO employees have a CSIRO email address following the pattern "firstname.lastname@csiro.au", we extract a list of candidates with email addresses matching this pattern from text. The candidates' full names, other names, and other email addresses are also extracted from text using regular expression patterns, and grouped with their CSIRO email addresses using identity matching techniques. Advanced named entity recognition techniques are used for generating variations of people's names. People's full names, name variations, email addresses, user IDs etc. are matched using the Aho-Corasick algorithm. The total number of candidates is 3,483 with 808,148 occurrences in the dataset.

The occurrences of candidates follow the power law distribution, where a small number of candidates have a very large number of occurrences, and a majority of candidates have a small number of occurrences.

All experiments make use of the Indri search engine (http://www.lemurproject.org/indri/). Documents are stemmed using the Porter stemmer, but not stopped at index time. Instead, stopping is done at query time using a standard list of 421 stopwords.

We used the TREC2007 Enterprise Track 50 search topics for both ad hoc retrieval and expert finding on the CSIRO dataset.

For example, one search topic is:

<title>climate change</title>
<narrative>
CSIRO has recently provided the most authoritative assessments of the likely impact of global climate change for Australia. CSIRO conducts research into the changing climate, along with the social and environmental impacts of climate change. Find a number of relevant pages and experts on climate change, respectively.
</narrative>

The goal of ad hoc retrieval is to find documents relevant to "climate change", and expert finding is to find people with expertise on the study of "climate change" in CSIRO, respectively.

We used only the query title part for both ad hoc retrieval and expert finding, respectively.

TREC2007 Enterprise provides answers to these 50 topics as ground truth on both ad hoc and expert retrieval. We use the ground truth as the gold standard in our evaluation.

We report MAP (mean average precision), the main performance measure for both ad hoc retrieval and expert finding in TREC. AP (average precision) for a topic emphasizes returning more relevant documents earlier, and is average of precisions after each of the relevant documents returned. MAP is the mean of the APs for the 50 topics. Where stated, we tested statistical significance with t tests (one-tail critical values for significance levels α=0.05).

Firstly, we study the effect of background smoothing in ad hoc retrieval and expert finding, i.e., by adjusting μ in Equation 5. Previous research on language model shows that value of μ around 2000 in Equation 5 leads to good ad hoc retrieval performance [18]. We set the value of μ as 1, 100, 500, 1000, and 2000, and get the MAP for ad hoc retrieval on the CSIRO dataset, respectively, as shown in Table 1. We can see that when μ increases from 1 to a value above 100, the MAP dramatically improves. The improvement in MAP from μ =100 to μ =500 is still statistically significant. When μ is 1000 or above, MAP keeps stable.

Now we will study how the value of μ influences expert finding. We select five representative window sizes, i.e., 50, 100, 300, 500, and 800. The results are reported in Table 2.

We can see from Table 2 that more effective document retrieval often does not lead to more effective expert finding, e.g., the improvement in terms of ad hoc retrieval from μ =1 to μ =100 is dramatic, however, the expert finding performance for all window sizes does not improve or even degrades for most window sizes, however, none of these improvement or degradation in performance are statistically significant. For window size of 50 and 100, the best performance is achieved when μ =1, i.e., when the effect of background model is small. For window size of 300, large value of μ leads improvement in expert finding, but not statistically significant. For window size of 500 and 800, expert finding performance keeps stable for all five window sizes, respectively.

We summarize from Table 1 and 2 that the introduction of background smoothing in language model is effective for ad hoc retrieval, but not effective for expert finding. The reason might be that expert finding based on co-occurrences of experts and query terms is insensitive to improvement in ad hoc retrieval by introducing background information.

Table 1. MAP of ad hoc retrieval for different μ, and percentage of improvement over the MAP for μ =1.

μ	1	100	500	1000	2000
MAP	0.287	0.333 +16%	0.383 +33%	0.389 +35%	0.380 +32%

Table 2. MAP of expert finding for different window size and μ, and percentage of improvement over the MAP for μ =1 and a certain window size, respectively.

Window	50	100	300	500	800
μ =1	0.418	0.428	0.367	0.366	0.364
μ =100	0.409 -2.1%	0.418 -2.3%	0.366 -0.3%	0.367 +0.3%	0.362 -0.5%
μ =500	0.402 -3.8%	0.414 -3.3%	0.381 +3.8%	0.366 +0%	0.363 -0.3%
μ =1000	0.402 -3.8%	0.415 -3.0%	0.379 +3.3%	0.365 -0.3%	0.361 -0.8%
μ =2000	0.408 -2.4%	0.415 -3.0%	0.379 +3.3%	0.374 +2.2%	0.358 -1.6%

To study the effect of anchor texts and indegree in ad hoc retrieval and expert finding, we fixed μ =1000, and experimented with different configurations of λ_t and λ_a in Equation 4, and indegree in Equation 2.

We varied λ_a from 0.0 to 1.0 in steps of 0.1. The curve for MAP is shown in Figure 1. In Figure 1, when the weight of anchor text increases, the MAP of ad hoc retrieval for both with and without indegree decreases. When λ_a =0.4 or above, the decrease of MAP from the document content only based model, i.e., λ_a =0.0, is statistically significant.

In Figure 1, the introduction of indegree helps improve the performance of purely document content based, anchor text and document content based mixture model, and purely anchor text based models, respectively. However these increases in MAP are not statistically significant.

Our baseline for expert finding is a basic single window based language model where anchor texts are not used, i.e., λ_t=1.0, λ_a=0.0 and μ =1000.

To study the effect of indegree and anchor texts in expert finding, we varied λ_a from 0.0 to 1.0 in steps of 0.1 without/with indegree in an integrated model for expert finding.

In order to easily compare a basic single window based expert finding approach and the basic single window enhanced by anchor texts and indegree, we plot the percentage of gain on MAP (pgMAP) of these enhanced models for different window sizes in Figure 2. We define pgMAP as

$$pgMAP_W = \frac{MAP_{EM,W} - MAP_{BM,W}}{MAP_{BM,W}},$$

where $MAP_{EM,W}$ is the MAP for the enhanced model of a certain window size W, and $MAP_{BM,W}$ is the MAP for the basic model of window W.

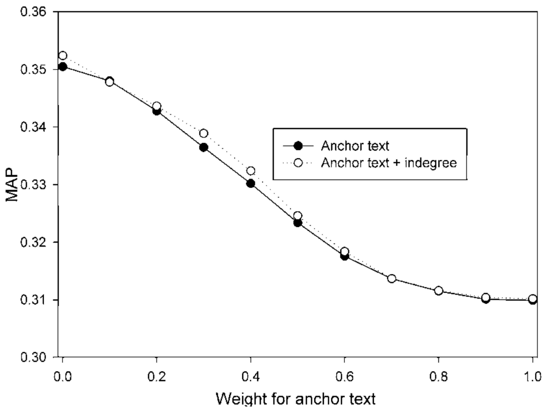

Figure 1: MAP of document retrieval for different anchor text weights with/without indegree

In Figure 2a and 2b, when the window size is 480 or above, all the anchor text enhanced models perform statistically significantly better than the basic models for both with and without indegree, respectively. Comparing Figure 2a and 2b, we can see that indegree has further helped improve the MAP of the anchor text alone enhanced models for window size 480 or above.

From Figure 1 and 2, we can see that anchor text does not help ad hoc retrieval, and hurt ad hoc retrieval when weighted high in a language model approach, and indegree only slightly helps ad hoc retrieval. However, both anchor text and indegree help expert finding, and anchor text helps expert finding with statistical significance when window size is above 480.

We think that our observation about anchor text and indegree in ad hoc retrieval and expert finding is due to the nature of expert finding in an organizational environment.

The TREC2007 search topics were created by CSIRO science communicators and these topics reflect typical enquiries from the media. These topics are generally well known research areas inside the CSIRO, and potential experts are often key contacts mentioned in authoritative documents on these topics, such as a number of experts on "climate change" are listed on the CSIRO climate research homepage at (http://www.cmar.csiro.au/research/climate.html).

These authoritative documents typically have more links from other pages and keywords on the topic often occur in anchor texts pointing to these documents.

By integrating anchor text and indegree in ad hoc retrieval, highly authoritative documents on a topic tend to be judged more relevant than ordinary documents. Therefore, people appearing in these highly authoritative documents will be given more credit in expert finding. Since these people are more likely to be experts on the topic, expert finding is improved.

On the other hand, in ad hoc retrieval, documents are largely judged as relevant or not regardless of their authoritativeness, and the relevant set includes both authoritative and ordinary documents, such as a review on "climate change" (http://www.csiro.au/files/files/pleo.pdf) is judged as relevant to the topic "climate change" where experts on the topic are not mentioned. The incorporation of anchor text and indegree may introduce more noise than useful information in ad hoc retrieval, therefore, leads to no improvement or even hurt the performance of ad hoc retrieval.

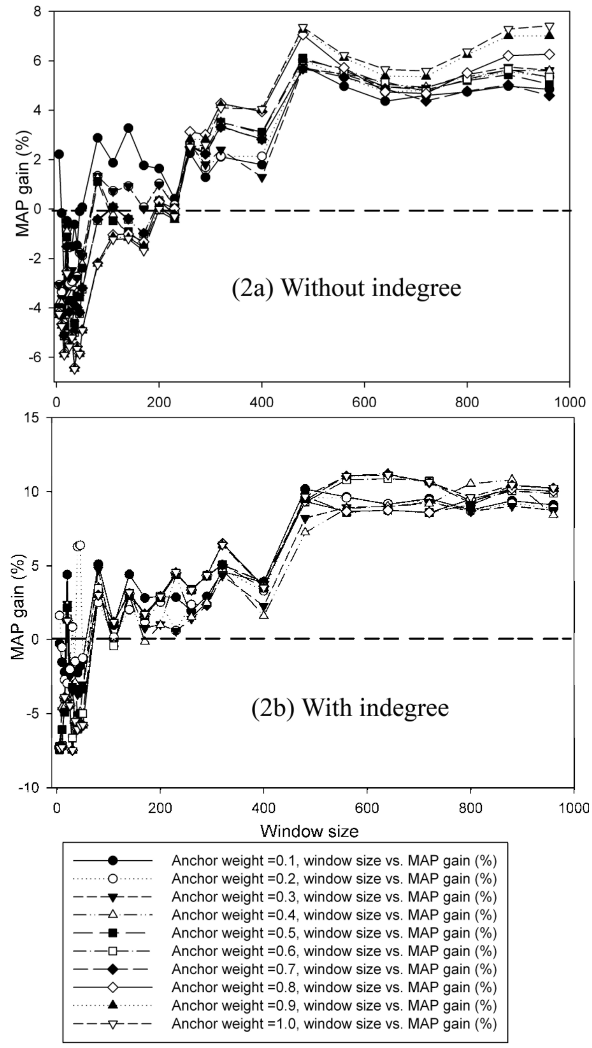

Figure 2: When anchor text takes different weights for with/without indegree, percentage of gain on MAP (pgMAP) for different window sizes

5. CONCLUSIONS AND FUTURE WORK

Expert finding is an emerging research topic in information retrieval which has attracted considerable attention from both industry and academic communities. A two-stage language model based approach has been applied to expert finding with good performance. Ad hoc retrieval plays an important role in expert finding by locating documents relevant to an expertise request in the form of a query. In this paper, we studied the relationship between ad hoc retrieval and expert finding via three parameters, namely, a background smoothing parameter in language model, and anchor text and indegree of documents. Our experiments on the TREC 2007 Enterprise Track CSIRO dataset show that improvement in ad hoc retrieval does not necessarily lead to improvement in expert finding, and the nature of expert finding in an organizational environment also has effect in the expert finding performance. In summary, our findings are as follows.

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. We think that the reason might be that expert finding focuses more on co-occurrences of experts and query terms, therefore, is insensitive to improvement in ad hoc retrieval by background smoothing.

Secondly, anchor text does not help ad hoc retrieval, and hurts ad hoc retrieval when weighted high in ad hoc retrieval, and indegree only slight helps ad hoc retrieval. However, both anchor text and indegree help expert finding, and anchor text helps expert finding with statistical significance when window size is above 480. The reason is that integration of anchor text and indegree will favor more authoritative documents on a topic than ordinary documents in ad hoc retrieval. Since people appearing in authoritative documents are more likely to be experts than those appearing in ordinary documents, anchor text and indegree help expert finding. On the other hand, in ad hoc 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 ad hoc retrieval.

Our research gives insight into how to design an effective approach for both ad hoc retrieval and expert finding for enterprises, e.g., parameters correlating with document authority such as anchor text and indegree should be emphasized in expert finding but not for ad hoc retrieval, however, background smoothing should be emphasized in ad hoc retrieval but not in expert finding.

In our future work, we will carry out evaluations on other datasets such as the upcoming TREC2008 Enterprise Track test collection to verify our findings. We will study whether PageRank has the similar effect of indegree in ad hoc retrieval and expert finding, since PagerRank correlates with indegree [16]. We will also study the effect of other document features such as document internal structure in ad hoc retrieval and expert finding.

6. REFERENCES

[1] Bailey, P., Craswell, N., de Vries, A. P., Soboroff, I. (2007) Overview of the TREC 2007 Enterprise Track (DRAFT). In Proc. of TREC 2007 Notebook.

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

[3] 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.

[4] Cheng, T., Yan, X., and Chang, K. C-C (2007) EntityRank: Searching Entities Directly and Holistically. In Proc. of VLDB 2007: 387-398

[5] Craswell, N., and Hawking, D. (2005) Overview of the TREC-2004 Web Track. In Proc. of TREC 2004.

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

[7] Craswell, N., de Vries, A.P., Soboroff, I. (2006) Overview of the TREC-2005 Enterprise Track. In Proc. of TREC 2005.

[8] Eiron, N., and McCurley, K. S. (2003) Analysis of anchor text for web search. In Proc. of SIGIR 2003: 459-460.

[9] Fang, H., and Zhai, C. (2007) Probabilistic Models for Expert Finding. In Proc. of ECIR 2007: 418-430

[10] Fu, Y., Xiang, R., Zhang, M., Liu, Y., and Ma, S. (2006) A PDD-Based Searching Approach for Expert Finding in Intranet Information Management. In Proc. of AIRS 2006: 43-53

[11] Petkova, D., and Croft, W. B. (2006) Hierarchical Language Models for Expert Finding in Enterprise Corpora. In Proc. of IEEE International Conf. on Tools with Artificial Intelligence: 599-608

[12] Petkova, D., and Croft, W. B. (2007) Proximity-based document representation for named entity retrieval. In Proc. of CIKM: 731-740.

[13] Ponte, J.M., and Croft, W.B. (1998) A Language Modeling Approach to Information Retrieval. In Proc. of SIGIR 1998: 275-281.

[14] Serdyukov, P., and Hiemstra, D. (2008) Modeling documents as mixtures of persons for expert finding. In Proc. of ECIR 2008.

[15] Soboroff, I., de Vries, A.P. and Craswell, N. (2007) Overview of the TREC 2006 Enterprise Track. In TREC 2006.

[16] 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.

[17] Yimam-Seid, D. and Kobsa, A. (2003) Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Journal of Organizational Computing & Electronic Commerce 13 (1).

[18] Zhai, C., and Lafferty, J. (2004) A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS), Vol. 22, No. 2: 179-214.

[19] Zhu, J., Gonçalves, A.L., Uren, V.S., Motta, E., and Pacheco, R., Eisenstadt, M., and Song, D. (2007) Mining Web Data for Competency Management. Relation Discovery from Web Data for Competency Management. Web Intelligence and Agent Systems: An International Journal, Vol. 5, No. 4, 405-417.

[20] 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 TREC 2006.

[21] Zhu, J., Song, D., Rüger, S., and Huang, J. (2008) Modeling Document Features for Expert Finding. To appear in Proc. of ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California.