# A Probabilistic Model for Fine-Grained Expert Search

**Shenghua Bao[1], Huizhong Duan[1], Qi Zhou[1], Miao Xiong[1], Yunbo Cao[1,2], Yong Yu[1]**

[1]Shanghai Jiao Tong University,
Shanghai, China, 200240
{shhbao,summer,jackson,xiongmiao,yyu}
@apex.sjtu.edu.cn

[2]Microsoft Research Asia
Beijing, China, 100080
yunbo.cao@microsoft.com

## Abstract

Expert search, in which given a query a ranked list of experts instead of documents is returned, has been intensively studied recently due to its importance in facilitating the needs of both information access and knowledge discovery. Many approaches have been proposed, including metadata extraction, expert profile building, and formal model generation. However, all of them conduct expert search with a coarse-grained approach. With these, further improvements on expert search are hard to achieve. In this paper, we propose conducting expert search with a fine-grained approach. Specifically, we utilize more specific evidences existing in the documents. An evidence-oriented probabilistic model for expert search and a method for the implementation are proposed. Experimental results show that the proposed model and the implementation are highly effective.

## 1  Introduction

Nowadays, team work plays a more important role than ever in problem solving. For instance, within an enterprise, people handle new problems usually by leveraging the knowledge of experienced colleagues. Similarly, within research communities, novices step into a new research area often by learning from well-established researchers in the research area. All these scenarios involve asking the questions like "who is an expert on X?" or "who knows about X?" Such questions, which cannot be answered easily through traditional document search, raise a new requirement of searching people with certain expertise.

To meet that requirement, a new task, called *expert search*, has been proposed and studied intensively. For example, TREC 2005, 2006, and 2007

provide the task of *expert search* within the enterprise track. In the TREC setting, e*xpert search* is defined as: given a query, a ranked list of experts is returned. In this paper, we engage our study in the same setting.

Many approaches to expert search have been proposed by the participants of TREC and other researchers. These approaches include metadata extraction (Cao et al., 2005), expert profile building (Craswell, 2001, Fu et al., 2007), data fusion (Maconald and Ounis, 2006), query expansion (Macdonald and Ounis, 2007), hierarchical language model (Petkova and Croft, 2006), and formal model generation (Balog et al., 2006; Fang et al., 2006). However, all of them conduct expert search with what we call a coarse-grained approach. The discovering and use of evidence for expert locating is carried out under a grain of document. With it, further improvements on expert search are hard to achieve. This is because different blocks (or segments) of electronic documents usually present different functions and qualities and thus different impacts for expert locating.

In contrast, this paper is concerned with proposing a probabilistic model for fine-grained expert search. In fine-grained expert search, we are to extract and use evidence of expert search (usually blocks of documents) directly. Thus, the proposed probabilistic model incorporates evidence of expert search explicitly as a part of it. A piece of *fine-grained evidence* is formally defined as a quadruple, <*topic*, *person*, *relation*, *document*>, which denotes the fact that a *topic* and a *person*, with a certain *relation* between them, are found in a specific *document*. The intuition behind the quadruple is that a query may be matched with phrases in various forms (denoted as *topic* here) and an *expert candidate* may appear with various name masks (denoted as *person* here), e.g., full name, email, or abbreviated names. Given a *topic* and *person*, *relation* type is used to measure their closeness and

*document* serves as a context indicating whether it is good *evidence*.

Our proposed model for fine-grained expert search results in an implementation of two stages.

 1) Evidence Extraction: document segments in various granularities are identified and evidences are extracted from them. For example, we can have segments in which an expert candidate and a queried topic co-occur within a same section of *document-001*: "…later, *Berners-Lee* describes a *semantic web search engine* experience…" As the result, we can extract an evidence by using same-section relation, i.e., <*semantic web search engine*, *Berners-Lee*, *same-section*, *document-001*>.

2) Evidence Quality Evaluation: the quality (or reliability) of evidence is evaluated. The quality of a quadruple of evidence consists of four aspects, namely *topic-matching quality*, *person-name-matching quality*, *relation quality*, and *document quality*. If we regard evidence as link of expert candidate and queried topic, the four aspects will correspond to the strength of the link to query, the strength of the link to expert candidate, the type of the link, and the document context of the link respectively.

All the *evidence*s with their scores of *quality* are merged together to generate a single score for each expert candidate with regard to a given query. We empirically evaluate our proposed model and implementation on the W3C corpus which is used in the expert search task at TREC 2005 and 2006. Experimental results show that both explored evidences and evaluation of evidence quality can improve the expert search significantly. Compared with existing state-of-the-art expert search methods, the probabilistic model for fine-grained expert search shows promising improvement.

The rest of the paper is organized as follows. Section 2 surveys existing studies on expert search. Section 3 and Section 4 present the proposed probabilistic model and its implementation, respectively. Section 5 gives the empirical evaluation. Finally, Section 6 concludes the work.

## 2   Related Work

### 2.1   Expert Search Systems

One setting for automatic expert search is to assume that data from specific resources are available. For example, Expertise Recommender (Kautz et al., 1996), Expertise Browser (Mockus and Herbsleb, 2002) and the system in (McDonald and Ackerman, 1998) make use of log data in software development systems to find experts. Yet another approach is to mine expert and expertise from email communications (Campbell et al., 2003; Dom et al. 2003; Sihn and Heeren, 2001).

Searching expert from general documents has also been studied (Davenport and Prusak, 1998; Mattox et al., 1999; Hertzum and Pejtersen, 2000). P@NOPTIC employs what is referred to as the 'profile-based' approach in searching for experts (Craswell et al., 2001). Expert/Expert-Locating (EEL) system (Steer and Lochbaum, 1988) uses the same approach in searching for expert groups. DEMOIR (Yimam, 1996) enhances the profile-based approach by separating co-occurrences into different types. In essence, the profile-based approach utilizes the co-occurrences between query words and people within documents.

### 2.2   Expert Search at TREC

A task on expert search was organized within the enterprise track at TREC 2005, 2006 and 2007 (Craswell et al., 2005; Soboroff et al., 2006; Bailey et al., 2007).

Many approaches have been proposed for tackling the expert search task within the TREC track. Cao et al. (2005) propose a two-stage model with a set of extracted metadata. Balog et al. (2006) compare two generative models for expert search. Fang et al. (2006) further extend their generative model by introducing the prior of expert distribution and relevance feedback. Petkova and Croft (2006) further extend the profile based method by using a hierarchical language model. Macdonald and Ounis (2006) investigate the effectiveness of the voting approach and the associated data fusion techniques. However, such models are conducted in a coarse-grain scope of document as discussed before. In contrast, our study focuses on proposing a model for conducting expert search in a fine-grain scope of evidence (local context).

## 3   Fine-grained Expert Search

Our research is to investigate a direct use of the local contexts for expert search. We call each local context of such kind as *fine-grained evidence*.

In this work, a fine-grained evidence is formally defined as a quadruple, <*topic, person, relation,*

*document>*. Such a quadruple denotes that a *topic* and a *person* occurrence, with a certain *relation* between them, are found in a specific *document*.

Recall that *topic* is different from *query*. For example, given a query *"semantic web coordination"*, the corresponding *topic* may be either *"semantic web"* or *"web coordination"*. Similarly, *person* here is different from *expert candidate*. E.g, given an *expert candidate "Ritu Raj Tiwari"*, the matched *person* may be *"Ritu Raj Tiwari"*, *"Tiwari"*, or *"RRT"* etc. Although both the topics and persons may not match the query and expert candidate exactly, they do have certain indication on the connection of query *"semantic web coordination"* and expert *"Ritu Raj Tiwari"*.

### 3.1 Evidence-Oriented Expert Search Model

We conduct fine-grained expert search by incorporating evidence of local context *explicitly* in a probabilistic model which we call an *evidence-oriented expert search model*. Given a query $q$, the probability of a candidate $c$ being an expert (or knowing something about $q$) is estimated as

$$P(c \mid q) = \sum_e P(c, e \mid q)$$
$$= \sum_e P(c \mid e, q) P(e \mid q) \tag{1}$$

where $e$ denotes a quadruple of *evidence*.

Using the relaxation that the probability of $c$ is independent of a query $q$ given an evidence $e$, we can reduce Equation (1) as,

$$P(c \mid q) = \sum_e P(c \mid e) P(e \mid q) \cdot \tag{2}$$

Compared to previous work, our model conducts expert search with a new way in which local contexts of *evidence* are used to bridge a query $q$ and an expert candidate $c$. The new way enables the expert search system to explore various local contexts in a precise manner.

In the following sub-sections, we will detail two sub-models: the *expert matching model* $P(c|e)$ and the *evidence matching model* $P(e|q)$.

### 3.2 Expert Matching Model

We expand the evidence $e$ as quadruple *<topic, people, relation, document>* (*<t, p, r, d>* for short) for expert matching. Given a set of related evidences, we assume that the generation of an expert candidate $c$ is independent with topic $t$ and omit it

in expert matching. Therefore, we simplify the expert matching formula as below:

$$P(c \mid e) = P(c \mid p, r, d) = P(c \mid p) P(p \mid r, d), \tag{3}$$

where $P(c|p)$ depends on how an expert candidate $c$ matches to a person occurrence $p$ (e.g. full name or email of a person). The different ways of matching an expert candidate $c$ with a person occurrence $p$ results in varied qualities. $P(c|p)$ represents the quality. $P(p|r,d)$ expresses the probability of an occurrence $p$ given a relation $r$ and a document $d$. $P(p|r,d)$ is estimated in MLE as,

$$P(p \mid r, d) = \frac{freq(p, r, d)}{L(r, d)}, \tag{4}$$

where $freq(p,r,d)$ is the frequency of person $p$ matched by relation $r$ in document $d$, and $L(r, d)$ is the frequency of all the persons matched by relation $r$ in $d$. This estimation can further be smoothed by using the evidence collection as follows:

$$P_S(p \mid r, d) = \mu P(p \mid r, d) + (1 - \mu) \sum_{d \in D} \frac{P(p \mid r, d')}{\mid D \mid}, \tag{5}$$

where $D$ denotes the whole document collection. $|D|$ is the total number of documents.

We use Dirichlet prior in smoothing of parameter $\mu$:

$$\mu = \frac{L(r, d)}{L(r, d) + K}, \tag{6}$$

where $K$ is the average frequency of all the experts in the collection.

### 3.3 Evidence Matching Model

By expanding the evidence $e$ and employing independence assumption, we have the following formula for evidence matching:

$$P(e \mid q) = P(t, p, r, d \mid q)$$
$$= P(t \mid q) P(p \mid q) P(r \mid q) P(d \mid q) \cdot \tag{7}$$

In the following, we are to explain what these four terms represent and how they can be estimated.

The first term $P(t|q)$ represents the probability that a query $q$ matches to a topic $t$ in evidence. Recall that a query $q$ may match a topic $t$ in various ways, not necessarily being identical to $t$. For example, both topic *"semantic web"* and *"semantic web search engine"* can match the query *"semantic web search engine"*. The probability is defined as

$$P(t\,|\,q) \propto P\big(type(t,q)\big), \qquad (8)$$

where *type(t, q)* represents the way that *q* matches to *t,* e.g., *phrase matching*. Different matching methods are associated with different probabilities.

The second term *P(p|q)* represents the probability that a person *p* is generated from a query *q*. The probability is further approximated by the prior probability of *p,*

$$P(p\,|\,q) \propto P(p)\,. \qquad (9)$$

The prior probability can be estimated by MLE, i.e., the ratio of total occurrences of person *p* in the collection.

The third term represents the probability that a relation *r* is generated from a query *q*. Here, we approximate the probability as

$$P(r\,|\,q) \propto P(type(r))\,, \qquad (10)$$

where *type(r)* represents the way *r* connecting query and expert. *P(type(r))* represents the reliability of relation type of *r*.

Following the Bayes rule, the last term can be transformed as

$$P(d\,|\,q) = \frac{P(q\,|\,d)P(d)}{P(q)} \propto P(q\,|\,d)P(d)\,, \qquad (11)$$

where priority distribution *P(d)* can be estimated based on static rank, e.g., PageRank (Brin and Page, 1998). *P(q|d)* can be estimated by using a standard language model for IR (Ponte and Croft, 1998).

In summary, Equation (7) is converted to

$$P(e\,|\,q) \propto P\big(type(t,q)\big)P(p)P(type(r))P(q\,|\,d)P(d)\,. \qquad (12)$$

### 3.4    Evidence Merging

We assume that the ranking score of an expert can be acquired by summing up together all scores of the supporting evidences. Thus we calculate experts' scores by aggregating the scores from all evidences as in Equation (1).

## 4    Implementation

The implementation of the proposed model consists of two stages, namely *evidence extraction* and *evidence quality evaluation*.

### 4.1    Evidence Extraction

Recall that we define an evidence for expert search as a quadruple *<topic, person, relation, document>*. The evidence extraction covers the extraction of the first three elements, namely *person identification*, *topic discovering* and *relation extraction*.

#### 4.1.1    Person Identification

The occurrences of an expert can be in various forms, such as name and email address. We call each type of form an *expert mask*. Table 1 provides a statistic on various *masks* on the basis of W3C corpus. In Table 1, *rate* is the proportion of the person occurrences with relevant masks to the person occurrences with any of the masks, and *ambiguity* is defined as the probability that a mask is shared by more than one expert.

| Mask | Rate/Ambiguity | Sample |
|---|---|---|
| Full Name($N_F$) | 48.2% / 0.0000 | Ritu Raj Tiwari |
| Email Name($N_E$) | 20.1% / 0.0000 | rtiwari@nuance.com |
| Combined Name ($N_C$) | 4.2% /0.3992 | Tiwari, Ritu R; R R Tiwari |
| Abbr. Name($N_A$) | 21.2% / 0.4890 | Ritu Raj ; Ritu |
| Short Name($N_S$) | 0.7% / 0.6396 | RRT |
| Alias, new email ($N_{AE}$) | 7% / 0.4600 | Ritiwari rtiwari@hotmail.com |

Table 1. Various masks and their ambiguity

| | |
|---|---|
| 1) | Every occurrence of a candidate's *email address* is normalized to the appropriate *candidate_id*. |
| 2) | Every occurrence of a candidate's *full_name* is normalized to the appropriate *candidate_id* if there is no ambiguity; otherwise, the occurrence is normalized to the *candidate_id* of the most frequent candidate with that *full_name*. |
| 3) | Every occurrence of *combined name*, *abbreviated name*, and *email alias* is normalized to the appropriate *candidate_id* if there is no ambiguity; otherwise, the occurrence may be normalized to the *candidate_id* of a candidate whose full name also appears in the document. |
| 4) | All the personal occurrences other than those covered by Heuristic 1) ~ 3) are ignored. |

Table 2. Heuristic rules for expert extraction

As Table 1 demonstrates, it is not an easy task to identify all the *mask*s with regards to an expert. On one hand, the extraction of *full name* and *email address* is straightforward but suffers from low coverage. On the other hand, the extraction of

*combined name* and *abbreviated name* can complement the coverage, while needs handling of ambiguity.

Table 2 provides the heuristic rules that we use for expert identification. In the step 2) and 3), the rules use *frequency* and *context discourse* for resolving ambiguities respectively. With *frequency,* each expert candidate actually is assigned a prior probability. With *context discourse,* we utilize the intuition that person names appearing similar in a document usually refers to the same person.

### 4.1.2 Topic Discovering

A queried topic can occur within documents in various forms, too. We use a set of query processing techniques to handle the issue. After the processing, a set of topics transformed from an original query will be obtained and then be used in the search for experts. Table 3 shows five forms of topic discovering from a given query.

| Forms | Description | Sample |
|---|---|---|
| Phrase Match($Q_P$) | The exact match with original query given by users | "*semantic web search engine*" |
| Bi-gram Match($Q_B$) | A set of matches formed by extracting bi-gram of words in the original query | "*semantic web*" "*search engine*" |
| Proximity Match($Q_{PR}$) | Each query term appears as a neighborhood within a window of specified size | "*semantic web enhanced search engine*" |
| Fuzzy Match($Q_F$) | A set of matches, each of which resembles the original query in appearance. | "*sementic web seerch engine*" |
| Stemmed Match($Q_S$) | A match formed by stemming the original query. | "*sementic web seerch engin*" |

Table 3. Discovered topics from query "semantic web search engine"

### 4.1.3 Relation Extraction

We focus on extracting relations between topics and expert candidates within a span of a document. To make the extraction easier, we partition a document into a pre-defined layout. Figure 1 provides a template in Backus–Naur form. Figure 2 provides a practical use of the template.

Note that we are not restricting the use of the template only for certain corpus. Actually the template can be applied to many kinds of documents. For example, for web pages, we can construct the <Title> from either the 'title' metadata or the con-

tent of web pages (Hu et al., 2006). As for e-mail, we can use the 'subject' field as the <Title>.

```
<Document> ::= {<Title>| <Author> |<Body>}
<Title> ::= {<Element>}
<Author> ::= {<Expert>}
<Body> ::= {<Section>}
<Section> ::= {<Section Title><Section Body>}
<SectionTitle> ::= {<Element>}
<SectionBody> ::= {<Section>}|{< Element >}
<Element> ::= <Term>| <Expert>
```

Figure. 1. A template of document layout



Figure 2. An example use of the layout template

With the layout of partitioned documents, we can then explore many types of relations among different blocks. In this paper, we demonstrate the use of five types of relations by extending the study in (Cao et al., 2005).

**Section Relation ($R_S$)**: The queried topic and the expert candidate occur in the same *<Section>*.

**Windowed Section Relation ($R_{WS}$)**: The queried topic and the expert candidate occur within a fixed window of a *<Section>*. In our experiment, we used a window of 200 words.

**Reference Section Relation ($R_{RS}$)**: Some *<Section>*s should be treated specially. For example, the *<Section>* consisting of reference information like a list of <book, author> can serve as a reliable source connecting a topic and an expert candidate. We call the relation appearing in a special type of *<Section>* a special reference section relation. It might be argued whether the use of special sections can be generalized. According to our survey, the special *<Section>*s can be found in various sites such as Wikipedia as well as W3C.

**Title-Author Relation ($R_{TA}$)**: The queried topic appears in the *<Title>* and the expert candidate appears in the *<Author>*.

**Section Title-Body Relation ($R_{STB}$)**: The queried topic and the expert candidate appear in the *<Section Title>* and *<Section Body>* of the same *<Section>*, respectively. Reversely, the queried topic and the expert candidate can appear in the *<Section Body>* and *<Section Title>* of a *<Section>*. The latter case is used to characterize the documents introducing certain expert or the expert introducing certain document.

Note that our model is not restricted to use these five relations. We use them only for the aim of demonstrating the flexibility and effectiveness of fine-grained expert search.

## 4.2 Evidence Quality Evaluation

In this section, we elaborate the mechanism used for evaluating the quality of evidence.

### 4.2.1 Topic-Matching Quality

In Section 4.1.2, we use five techniques in processing query matches, which yield five sets of match types for a given query. Obviously, the different query matches should be associated with different weights because they represent different qualities.

We further note that different bi-grams generated from the same query with the *bi-gram matching* method might also present different qualities. For example, both topic "*css test*" and "*test suite*" are the bi-gram matching for query "*css test suite*"; however, the former might be more informative. To model that, we use the number of returned documents to refine the query weight. The intuition behind that is similar to the thought of IDF popularly used in IR as we prefer to the distinctive bi-grams.

Taking into consideration the above two factors, we calculate the topic-matching quality $Q_t$ (corresponding to $P(type(t,q))$ in Equation (12) ) for the given query $q$ as

$$Q_t = W(type(t,q)) \frac{MIN_{t'}(df_{t'})}{df_t}, \qquad (13)$$

where $t$ means the discovered topic from a document and $type(t,q)$ is the matching type between topic $t$ and query $q$. $W(type(t,q))$ is the weight for a certain query type, $df_t$ is the number of returned documents matched by topic $t$. In our experiment, we use the 10 training topics of TREC2005 as our training data, and the best quality scores for *phrase match*, *bi-gram match*, *proximity match*, *fuzzy*

*match*, and *stemmed match* are 1, 0.01, 0.05, $10^{-8}$, and $10^{-4}$, respectively.

### 4.2.2 Person-Matching Quality

An expert candidate can occur in the documents in various ways. The most confident occurrence should be the ones in full name or email address. Others can include last name only, last name plus initial of first name, etc. Thus, the action of rejecting or accepting a person from his/her mask (the surface expression of a person in the text) is not simply a Boolean decision, but a probabilistic one with a reliability weight $Q_p$ (corresponding to $P(c|p)$ in Equation (3) ). Similarly, the best trained weights for *full name*, *email name*, *combined name*, a*bbreviated name*, *short name*, and *alias email* are set to 1, 1, 0.8, 0.2, 0.2, and 0.1, respectively.

### 4.2.3 Relation Type Quality

The relation quality consists of two factors. One factor is about the type of the relation. Different types of relations indicate different strength of the connection between expert candidates and queried topics. In our system, the *section title-body relation* is given the highest confidence. The other factor is about the degree of proximity between a query and an expert candidate. The intuition is that, the more distant are a query and an expert candidate within a relation, the looser the connection between them is. To include these two factors, the quality score $Q_r$ (corresponding to $P(type(r))$ in Equation (12) )of a relation $r$ is defined as:

$$Q_r = W_r \frac{C_r}{dis(p,t)+1}, \qquad (14)$$

where $W_r$ is the weight of relation type $r$, $dis(p, t)$ is the distance from the person occurrence $p$ to the queried topic $t$ and $C_r$ is a constant for normalization. Again, we optimize the $W_r$ based on the training topics, the best weights for *section relation*, *windowed section relation*, *reference section relation*, *title-author relation*, and *section title-body relation* are 1, 4, 10, 45, and 1000 respectively.

### 4.2.4 Document Quality

The quality of evidence also depends on the quality of the document, the context in which it is found. The *document context* can affect the credibility of the evidence in two ways:

**Static quality**: indicating the authority of a document. In our experiment, the static quality $Q_d$ (corresponding to $P(d)$ in Equation (12) ) is estimated by the PageRank, which is calculated using a standard iterative algorithm with a damping factor of 0.85 (Brin and Page, 1998).

**Dynamic quality**: by "dynamic", we mean the quality score varies for different queries $q$. We denote the dynamic quality as $Q_{DY}(d,q)$ (corresponding to $P(q|d)$ in Equation (12) ), which is actually the document relevance score returned by a standard language model for IR(Ponte and Croft, 1998).

## 5 Experimental Results

### 5.1 The Evaluation Data

In our experiment, we used the data set in the expert search task of enterprise search track at TREC 2005 and 2006. The document collection is a crawl of the public W3C sites in June 2004. The crawl comprises in total 331,307 web pages. In the following experiments, we used the training set of 10 topics of TREC 2005 for tuning the parameters aforementioned in Section 4.2, and used the test set of 50 topics of TREC 2005 and 49 topics of TREC 2006 as the evaluation data sets.

### 5.2 Evaluation Metrics

We used three measures in evaluation: *Mean average precision* (*MAP*), *R-precision* (*R-P*), and *Top N precision* (*P@N*). They are also the standard measures used in the expert search task of TREC.

### 5.3 Evidence Extraction

In the following experiments, we constructed the baseline by using the query matching methods of *phrase matching*, the expert matching methods of *full name matching and email matching*, and the relation of *section relation*. To show the contribution of each individual method for evidence extraction, we incrementally add the methods to the baseline method. In the following description, we will use '+' to denote applying new method on the previous setting.

### 5.3.1 Query Matching

Table 4 shows the results of expert search achieved by applying different methods of query matching.

$Q_B$, $Q_{PR}$, $Q_F$, and $Q_S$ denote bi-gram match, proximity match, fuzzy match, and stemmed match, respectively. The performance of the proposed model increases stably on MAP when new query matches are added incrementally. We also find that the introduction of $Q_F$ and $Q_S$ bring some drop on R-Precision and P@10. It is reasonable because both $Q_F$ and $Q_S$ bring high recall while affect the precision a bit. The overall relative improvement of using query matching compared to the baseline is presented in the row "Improv.". We performed t-tests on MAP. The p-values ($< 0.05$) are presented in the "T-test" row, which shows that the improvement is statistically significant.

| | TREC 2005 | | | TREC 2006 | | |
|---|---|---|---|---|---|---|
| | MAP | R-P | P@10 | MAP | R-P | P@10 |
| Baseline | 0.1840 | 0.2136 | 0.3060 | 0.3752 | 0.4585 | 0.5604 |
| +$Q_B$ | 0.1957 | 0.2438 | 0.3320 | 0.4140 | 0.4910 | 0.5799 |
| +$Q_{PR}$ | 0.2024 | **0.2501** | **0.3360** | 0.4530 | **0.5137** | **0.5922** |
| +$Q_F$,$Q_S$ | **0.2030** | **0.2501** | **0.3360** | **0.4580** | 0.5112 | 0.5901 |
| Improv. | 10.33% | 17.09% | 9.80% | 22.07% | 11.49% | 5.30% |
| T-test | 0.0084 | | | 0.0000 | | |

Table 4. The effects of query matching

### 5.3.2 Person Matching

For person matching, we considered four types of masks, namely *combined name* ($N_C$), *abbreviated name* ($N_A$), *short name* ($N_S$) and *alias and new email* ($N_{AE}$). Table 5 provides the results on person matching at TREC 2005 and 2006. The baseline is the best model achieved in previous section. It seems that there is little improvement on P@10 while an improvement of 6.21% and 14.00% is observed on MAP. This might be due to the fact that the matching method such as $N_C$ has a higher recall but lower precision.

| | TREC 2005 | | | TREC 2006 | | |
|---|---|---|---|---|---|---|
| | MAP | R-P | P@10 | MAP | R-P | P@10 |
| Baseline | 0.2030 | 0.2501 | 0.3360 | 0.4580 | 0.5112 | 0.5901 |
| +$N_C$ | 0.2056 | 0.2539 | **0.3463** | 0.4709 | 0.5152 | 0.5931 |
| +$N_A$ | 0.2106 | 0.2545 | 0.3400 | 0.5010 | 0.5181 | **0.6000** |
| +$N_S$ | 0.2111 | 0.2578 | 0.3400 | 0.5121 | 0.5192 | **0.6000** |
| +$N_{AE}$ | **0.2156** | **0.2591** | 0.3400 | **0.5221** | **0.5212** | **0.6000** |
| Improv. | 6.21% | 3.60% | 1.19% | 14.00% | 1.96% | 1.68% |
| T-test | 0.0064 | | | 0.0057 | | |

Table 5. The effects of person matching

### 5.3.3 Multiple Relations

For relation extraction, we experimentally demonstrated the use of each of the five relations proposed in Section 4.1.3, i.e., section relation ($R_S$), windowed section relation ($R_{WS}$), reference section relation ($R_{RS}$), title-author relation ($R_{TA}$), and section title-body relation ($R_{STB}$). We used the best model achieved in previous section as the baseline. From Table 6, we can see that the section title-body relation contributes the most to the improvement of the performance. By using all the discovered relations, a significant improvement of 19.94% and 8.35% is achieved.

| | TREC 2005 | | | TREC 2006 | | |
|---|---|---|---|---|---|---|
| | **MAP** | **R-P** | **P@10** | **MAP** | **R-P** | **P@10** |
| Baseline | 0.2156 | 0.2591 | 0.3400 | 0.5221 | 0.5212 | 0.6000 |
| $+R_{WS}$ | 0.2158 | 0.2633 | 0.3380 | 0.5255 | 0.5311 | 0.6082 |
| $+R_{RS}$ | 0.2160 | 0.2630 | 0.3380 | 0.5272 | 0.5314 | 0.6061 |
| $+R_{TA}$ | 0.2234 | 0.2634 | 0.3580 | 0.5354 | 0.5355 | 0.6245 |
| $+R_{STB}$ | **0.2586** | **0.3107** | **0.3740** | **0.5657** | **0.5669** | **0.6510** |
| Improv. | 19.94% | 19.91% | 10.00% | 8.35% | 8.77% | 8.50% |
| T-test | 0.0013 | | | 0.0043 | | |

Table 6. The effects of relation extraction

### 5.4 Evidence Quality

The performance of expert search can be further improved by considering the evidence quality. Table 7 shows the results by considering the differences in quality.

We evaluated two kinds of evidence quality: *context static quality* ($Q_d$) and *context dynamic quality* ($Q_{DY}$). Each of the evidence quality contributes about 1%-2% improvement for MAP. The improvement from the PageRank that we calculated from the corpus implies that the web scaled rank technique is also effective in the corpus of documents. Finally, we find a significant relative improvement of 6.13% and 2.86% on MAP by using evidence qualities.

| | TREC 2005 | | | TREC 2006 | | |
|---|---|---|---|---|---|---|
| | **MAP** | **R-P** | **P@10** | **MAP** | **R-P** | **P@10** |
| Baseline | 0.2586 | 0.3107 | 0.3740 | 0.5657 | 0.5669 | 0.6510 |
| $+Q_d$ | 0.2711 | 0.3188 | 0.3720 | 0.5900 | 0.5813 | 0.6796 |
| $+Q_{DY}$ | **0.2755** | **0.3252** | **0.3880** | **0.5943** | **0.5877** | **0.7061** |
| Improv. | 6.13% | 4.67% | 3.74% | 2.86% | 3.67% | 8.61% |
| T-test | 0.0360 | | | 0.0252 | | |

Table 7. The effects of using evidence quality

### 5.5 Comparison with Other Systems

In Table 8, we juxtapose the results of our probabilistic model for fine-grained expert search with automatic expert search systems from the TREC evaluation. The performance of our proposed model is rather encouraging, which achieved comparable results to the best automatic systems on the TREC 2005 and 2006.

| | | **MAP** | **R-prec** | **Prec@10** |
|---|---|---|---|---|
| Rank-1 System | TREC2005 | 0.2749 | 0.3330 | 0.4520 |
| | TREC2006[1] | 0.5947 | 0.5783 | 0.7041 |
| Our System | TREC2005 | 0.2755 | 0.3252 | 0.3880 |
| | TREC2006 | 0.5943 | 0.5877 | 0.7061 |

Table 8. Comparison with other systems

## 6 Conclusions

This paper proposed to conduct expert search using a fine-grained level of evidence. Specifically, quadruple evidence was formally defined and served as the basis of the proposed model. Different implementations of evidence extraction and evidence quality evaluation were also comprehensively studied. The main contributions are:

1.  The proposal of fine-grained expert search, which we believe to be a promising direction for exploring subtle aspects of evidence.
2.  The proposal of probabilistic model for fine-grained expert search. The model facilitates investigating the subtle aspects of evidence.
3.  The extensive evaluation of the proposed probabilistic model and its implementation on the TREC data set. The evaluation shows promising expert search results.

In future, we are to explore more domain independent evidences and evaluate the proposed model on the basis of the data from other domains.

---

[1] This system, where cluster-based re-ranking is used, is a variation of the fine-grained model proposed in this paper.

# References

Bailey, P., Soboroff , I., Craswell, N., and Vries A.P., Overview of the TREC 2007 Enterprise Track. In: *Proc. of TREC 2007*.

Balog, K., Azzopardi, L., and Rijke, M. D., 2006. Formal models for expert finding in enterprise corpora. In: *Proc. of SIGIR'06,*pp.43-50.

Brin, S. and Page, L., 1998. The anatomy of a rlarge-scale hypertextual Web search engine, *Computer Networks and ISDN Systems* (30), pp.107-117.

Campbell, C.S., Maglio, P., Cozzi, A. and Dom, B., 2003. Expertise identification using email communications. In: *Proc. of CIKM '03* pp.528–531.

Cao, Y., Liu, J., and Bao, S., and Li, H., 2005. Research on expert search at enterprise track of TREC 2005. In: *Proc. of TREC 2005*.

Craswell, N., Hawking, D., Vercoustre, A. M. and Wilkins, P., 2001. P@NOPTIC Expert: searching for experts not just for documents. In: *Proc. of Ausweb*'01.

Craswell, N., Vries, A.P., and Soboroff, I., 2005. Overview of the TREC 2005 Enterprise Track. In: *Proc. of TREC 2005*.

Davenport, T. H. and Prusak, L., 1998. Working Knowledge: how organizations manage what they know. *Howard Business*, *School Press*, Boston, MA.

Dom, B., Eiron, I., Cozzi A. and Yi, Z., 2003. Graph-based ranking algorithms for e-mail expertise analysis, In: *Proc. of SIGMOD'03 workshop on Research issues in data mining and knowledge discovery*.

Fang, H., Zhou, L., Zhai, C., 2006. Language models for expert finding-UIUC TREC 2006 Enterprise Track Experiments, In: *Proc. of TREC2006*.

Fu, Y., Xiang, R., Liu, Y., Zhang, M., Ma, S., 2007. A CDD-based Formal Model for Expert Finding. In *Proc. of CIKM 2007*.

Hertzum, M. and Pejtersen, A. M., 2000. The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing and Management,* 36(5), pp.761–778.

Hu, Y., Li, H., Cao, Y., Meyerzon, D. Teng, L., and Zheng, Q., 2006. Automatic extraction of titles from general documents using machine learning, *IPM*.

Kautz, H., Selman, B. and Milewski, A., 1996. Agent amplified communication. In: *Proc. of AAAI'96*, pp. 3–9.

Mattox, D., Maybury, M. and Morey, D., 1999. Enterprise expert and knowledge discovery. *Technical Report*.

McDonald, D. W. and Ackerman, M. S., 1998. Just Talk to Me: a field study of expertise location. In: *Proc. of CSCW'98*, pp.315-324.

Mockus, A. and Herbsleb, J.D*.,* 2002. Expertise Browser: a quantitative approach to identifying expertise, In: *Proc. of ICSE'02*.

Maconald, C. and Ounis, I., 2006. Voting for candidates: adapting data fusion techniques for an expert search task. In: *Proc. of CIKM'06*, pp.387-396.

Macdonald, C. and Ounis, I., 2007. Expertise Drift and Query Expansion in Expert Search. In *Proc. of CIKM 2007*.

Petkova, D., and Croft, W. B., 2006. Hierarchical language models for expert finding in enterprise corpora, In: *Proc.* of *ICTAI'06*, pp.599-608.

Ponte, J. and Croft, W., 1998. A language modeling approach to information retrieval, In: *Proc. of SIGIR'98*, pp.275-281.

Sihn, W. and Heeren F., 2001. Xpertfinder-expert finding within specified subject areas through analysis of e-mail communication. In: *Proc. of the 6th Annual Scientific conference on Web Technology*.

Soboroff, I., Vries, A.P., and Craswell, N., 2006. Overview of the TREC 2006 Enterprise Track. In: *Proc. of TREC 2006*.

Steer, L.A. and Lochbaum, K.E., 1988. An expert/expert locating system based on automatic representation of semantic structure, In: *Proc. of the 4th IEEE Conference on Artificial Intelligence Applications*.

Yimam, D., 1996. Expert finding systems for organizations: domain analysis and the DEMOIR approach. In: *ECSCW'99 workshop of beyond knowledge management: managing expertise*, pp. 276–283.