Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search

Yi Fang
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA
fangy@cs.purdue.edu

Luo Si
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA
lsi@cs.purdue.edu

Aditya P. Mathur
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA
apm@cs.purdue.edu

ABSTRACT

Generative models such as statistical language modeling have been widely studied in the task of expert search to model the relationship between experts and their expertise indicated in supporting documents. On the other hand, discriminative models have received little attention in expert search research, although they have been shown to outperform generative models in many other information retrieval and machine learning applications. In this paper, we propose a principled relevance-based discriminative learning framework for expert search and derive specific discriminative models from the framework. Compared with the state-of-the-art language models for expert search, the proposed research can naturally integrate various document evidence and document-candidate associations into a single model without extra modeling assumptions or effort. An extensive set of experiments have been conducted on two TREC Enterprise track corpora (i.e., W3C and CERC) to demonstrate the effectiveness and robustness of the proposed framework.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Design, Experimentation

Keywords

Expert search, enterprise search, discriminative models

1. INTRODUCTION

With vast amount of information available within large organizations, the key challenge is to harness existing knowledge and expertise in a timely and effective manner. In consequence, enterprise information retrieval systems are increasingly demanded to return people with specific knowledge and skills in response to a user's query. A class of vertical search engines known as expert finder have emerged for enterprise organizations.

As an important IR application, expert search (also known as expert finding) has received substantial attention in the IR research community. Rapid progress has been made in modeling and evaluation since the launch of TREC Enterprise Track in 2005 [12]. A notable observation is that probabilistic generative models have dominated the literature of expert search. In particular, many statistical language modeling techniques were proposed to model the relationship between a candidate expert and a query. These models usually characterize a generative process of how a query is generated from supporting documents of an expert. The key ingredient in these methods is to determine associations between people and documents because the associations are ambiguous in the TREC scenarios as well as in many realistic settings. Previous works have investigated different metrics or a combination of them to measure the associations, but the way of choosing or combining them is rather often heuristic and lacks of a clear justification. Furthermore, document evidence such as document or expert authority information, internal and external document structures, global evidence and so on is shown to be able to significantly improve expert retrieval performance, but to incorporate these features often requires many modeling assumptions and is often unwieldy.

On the other hand, discriminative models, another important class of probabilistic models with solid statistical foundation, are nearly absent in the research of expert search, especially on the TREC evaluations. In fact, discriminative models have been preferred over generative models in the recent past in many machine learning applications, partly because of their attractive theoretical properties. In the domain of IR, various discriminative models have also been applied to many retrieval problems (e.g., [23]). However, very limited research has been conducted to design discriminative models for expert search.

In this work, we present a relevance-based discriminative learning framework for expert search and derive specific discriminative models from the framework. Similar to some prominent language models, the proposed models aggregate document evidence and document-candidate associations through supporting documents. Unlike the language models, we directly model the conditional probability of relevance given a query and an expert. As a result, heterogeneous or even arbitrary features can be naturally included into a single model. The parameters associated with the features are automatically learned from training data. We

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

report an extensive set of experiments on two TREC corpora to evaluate the effectiveness and robustness of the proposed discriminative framework.

The next section discusses related work. Section 3 introduces the state-of-the-art generative language models for expert search. Section 4 presents our proposed approaches. In section 5, we discuss the advantages of discriminative models in the context of expert search. Section 6 explains our experimental methodology and Section 7 presents the experimental results. Section 8 concludes and points out some future work.

2. RELATED WORK

The early work on expert finding systems was initiated in the Knowledge Management community, usually in the form of yellow pages [9]. These systems relied on experts to judge and input their skills by themselves against a predefined set of keywords, and thus the task was time-consuming. More recent techniques locate experts in an automatic fashion. An overview of early automatic expert finding systems is provided in [36]. The task of expert search has received a significant amount of attention as the task had been included in the TREC Enterprise track from 2005 to 2008 [12, 32, 1, 7]. The TREC Enterprise tracks provided a common platform for researchers to empirically evaluate methods for expert search. They demonstrated the feasibility of expert search on heterogeneous data collections. In the TREC corpora, the relationship between documents and experts is ambiguous and thus to model the document-candidate associations is a key issue in expert search research.

Most of the recent work on expert search generally falls into two categories: profile-centric and document-centric approaches. Balog et al. [3] formalizes the two methods by proposing two generative language models. Their Model 1 directly models the knowledge of an expert from associated documents, which is equivalent to a profile-centric approach, and their Model 2 first locates documents on the topic and then finds the associated experts, which is a document-centric approach. It has been shown in [3] that Model 2 is generally more effective than Model 1 and since then it becomes one of the most prominent language models for expert search. In [8], a two-stage language model combining a document relevance and co-occurrence model is proposed, which is essentially equivalent to Model 2. An attempt to further improve their models is made by proposing a proximity-based document representation for incorporating sequential information in text [25]. There are many other generative probabilistic models proposed for expert finding. For example, Serdyukov and Hiemstra [30] propose an expert-centric language model. Fang and Zhai [14] derive two families of generative models by applying probability ranking principle. Probabilistic topic models are also proposed to simultaneously model the topical distribution of expertise evidence and experts [34].

Some alternative approaches to expert search exist beyond language modeling. One effective approach is to treat the problem of ranking experts as a voting problem based on data fusion techniques [21]. Eleven different voting strategies were proposed to aggregate over the documents associated to an expert. Another approach is to model the process of expert finding by probabilistic random walks on so-called expertise graphs [31]. Many other expert finding methods were proposed during TREC Enterprise tracks.

Besides the models, some researchers have shown that suitable features can help significantly boost the performance of expert finding. These features include document authority information such as the PageRank, indegree, and URL length [38], graph-based expert authority [10], internal document structures that indicate the experts' associations with the content of documents [6], non-local evidence [2], and the evidence that can be acquired outside of an enterprise [29]. Additional evidence can be integrated by identifying home pages of candidate experts and clustering relevant documents [20]. Proximity features that characterize the co-occurrence of query and expert mentions in the document are also shown indicative by the top runs in the TREC evaluations [16]. This led to several window-based approaches including [25, 4, 20].

On the other hand, the early work of applying discriminative models in IR can date back to the early 1980s in which the maximum entropy approach was investigated to get around term independence assumptions in probabilistic generative models [11]. More recently, Nallapati [23] compared the performance of the maximum entropy model and support vector machines with that of language modeling in ad hoc retrieval and homepage finding, and argued that SVMs are preferred over language models because of their ability to learn arbitrary features automatically. Furthermore, it has been shown that feature-based discriminative models can consistently and significantly outperform current state of the art retrieval models with the correct choice of features [22]. Discriminative models have received increasing attention in IR, as another related area, learning to rank for IR, sparked genuine interest among researchers in the community [18]. Most of the learning to rank models are discriminative in nature and they have been shown improvements over their generative counterparts in ad hoc retrieval. Benchmark data sets such as LETOR [19] are also available for research on learning to rank. Although valuable work has been done on discriminative models for ad hoc retrieval and other IR domains, very limited research has been conducted to design discriminative models for expert search. The only relevant work that we are aware of is [15], which addressed the issue of differentiating heterogeneous sources according to specific queries and experts by learning associated weights from data, but the work did not model document-candidate relationship nor address how to incorporate new document evidence, which are two key issues in expert search.

3. GENERATIVE MODELS

To predict a class C given an observation x, the desired choice of C is given by the conditional class probabilities $P(C|x)$. Depending on how to compute $P(C|x)$, the existing classification techniques can be broadly classified into two major categories: generative models and discriminative models. In a discriminative approach, a parametric model is introduced for $P(C|x)$, and the values of the parameters are inferred from a set of labeled training data. In contrast, the generative approach attempts to capture the manner in which an observation x is generated from given classes C by specifying a prior distribution $P(C)$ over classes and a class-conditional distribution $P(x|C)$ over the observation. The posterior $P(C|x)$ is obtained from Bayes' Theorem as

$$P(C|x) \propto P(x|C)P(C) \tag{1}$$

In the context of expert search, the task is to find out what is the probability of a candidate e being an expert given a query topic q. In other words, we want to know $P(e|q)$ in order to rank candidate e according to this probability. Similarly, by invoking Bayes' Theorem, we have:

$$P(e|q) \propto P(q|e)P(e) \tag{2}$$

where $P(e)$ is the prior probability of a candidate, which is generally assumed uniform. Thus, the key quantity to estimate in the generative models is the probability of a query given the candidate, $P(q|e)$. Many language modeling techniques are proposed to estimate this quantity. One of the most prominent and effective one was called document models (often referred as Model 2) [3] where documents act as a hidden variable in the process which accumulates expertise evidence. Formally, it is expressed as

$$P(q|e) = \sum_{t=1}^{n} P(q|d_t)P(d_t|e) \tag{3}$$

where $P(q|d_t)$ is the probability of the document d_t to generate the query q and can be calculated using a standard language model. $P(d_t|e)$ is the probability of association between the document d_t and the candidate e. n is the number of documents in the collection. Model 2 mimics the process one might use to find experts using a document retrieval system. Here, relevant documents are retrieved for the expertise requested, and they are used as evidence to indicate whether the associated candidates are experts. After aggregating all such evidence, the experts can be identified. As $P(q|d_t)$ is relatively easy to determine in language models, the key ingredient in this model (and also in many other language models for expert search) is to estimate the document-candidate associations: $P(d_t|e)$, or $P(e|d_t)$ if $P(d_t)$ is assumed to be uniform. $P(e|d_t)$ can be estimated by various methods. The simplest form is the boolean model where associations are binary decisions: $P(e|d_t) = 1$ if the candidate appears in the document; otherwise, $P(e|d_t) = 0$. More sophisticated methods are frequency based which consider the number of times that a candidate appears in the document. A set of heuristic combinations of all these metrics are also compared and investigated in [6].

4. DISCRIMINATIVE MODELS FOR EXPERT SEARCH

4.1 Discriminative Learning Framework for Expert Search

For the text-based retrieval, conventional relevance-based probabilistic models rank documents by sorting the conditional probability that each document would be judged relevant to the given query [17]. The underlying principle using probabilistic models for information retrieval is called probability ranking principle [26]. The Binary Independence Model (BIM) [27] is a realization of this principle. In the domain of expert search, the similar principle can be used where experts are ranked according to the descending order of the conditional probability of relevance given an expert and a query. Fang and Zhai [14] applied this principle in studying expert search problem. Both BIM and [14]'s models are generative and they use Bayes' theorem to reverse the original conditional probability.

We propose a discriminative learning framework to directly model the conditional probability of relevance by a parametric probability function. We cast expert search into a binary classification problem that treats the relevant query-expert pairs as positive data and irrelevant pairs as negative data. Formally, we use a relevance variable $r \in \{1, 0\}$ to denote whether two entities are relevant or not and thus the conditional probability of relevance $P_\theta(r|e, q)$ represents the extent to which the expert e is relevant to the query q. In our framework, $P_\theta(r|e, q)$ can take any function form with parameter θ that needs to estimate from training data. Based on different forms of P_θ, the resulting discriminative models are different. Given the relevance judgment r_{mk} for the training expert-query pair (e_k, q_m) which is assumed independently generated, the conditional likelihood L of the training data is as follows

$$L = \prod_{m}^{M}\prod_{k}^{K} P_\theta(r = 1|e_k, q_m)^{r_{mk}} P_\theta(r = 0|e_k, q_m)^{1-r_{mk}} \tag{4}$$

where M is the number of queries and K is the number of experts. The parameters can then be estimated by maximizing the following log likelihood function

$$\theta^* = \arg\max_\theta \sum_{m}^{M}\sum_{k}^{K}\Big(r_{mk}\log P_\theta(r = 1|e_k, q_m) + (1 - r_{mk})\log\big(1 - P_\theta(r = 1|e_k, q_m)\big)\Big) \tag{5}$$

The estimated parameters can then be plugged back in $P_\theta(r = 1|e_k, q_m)$. According to the probability ranking principle, the experts are presented to users in the descending order of $P_\theta(r = 1|e_k, q_m)$. In the next section, we propose a specific discriminative model by defining the form of $P_\theta(r = 1|e_k, q_m)$.

4.2 A Discriminative Model

According to the previous work, Model 2 turned out to be one of the most effective formal models for expert search. The success of the model lies in its effective process to collect expertise evidence from documents. Our discriminative model builds on the same process in which the supporting document d serves as a bridge to connect expert e and query q. Given a document d, whether e and q are relevant depends on two factors: document evidence and document-candidate associations. More specifically, we consider: 1) whether the document d is relevant to the query q; 2) whether the expert e is relevant to the document d. The final relevance decision for (e, q) is made by averaging over all the documents. Formally, this can be expressed as

$$P_\theta(r = 1|e, q) = \sum_{t=1}^{n} P(r_1 = 1|q, d_t)P(r_2 = 1|e, d_t)P(d_t) \tag{6}$$

where $P(r_1 = 1|q, d_t)$ allows us to model the probability that a document d_t matches a topic q, which indicates the document evidence. $P(r_2 = 1|e, d_t)$ allows us to model the probability that a supporting document d_t mentions a candidate e, which indicates the document-candidate associations. A document d_t with higher values on both probabilities would contribute more to the value of $P(r = 1|e, q)$. The prior probability of a document, $P(d_t)$, is generally assumed uniform (i.e., $P(d_t) = \frac{1}{n}$). We model both $P(r_1 = 1|q, d_t)$ and $P(r_2 = 1|e, d_t)$ by logistic functions on a linear combination of features. Formally, they are parameterized as follows:

$$P(r_1 = 1|q, d_t) = \sigma\big(\sum_{i=1}^{N_f} \alpha_i f_i(q, d_t)\big) \tag{7}$$

$$P(r_2 = 1|e, d_t) = \sigma\big(\sum_{j=1}^{N_g} \beta_j g_j(e, d_t)\big) \tag{8}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the standard logistic function. α_i is the weight for the i^{th} query-document feature $f_i(q, d_t)$ and β_j is the weight for the j^{th} document-candidate feature $g_j(e, d_t)$. Specifically, $f_i(q, d_t)$ is the document evidence such as document retrieval scores that indicates how relevant the document is to the query. $g_j(e, d_t)$ is the feature such as the boolean associations that describe the strength of associations between a document and a candidate. N_f denotes the number of document evidence features and N_g denotes the number of document-candidate association features. The weight parameters can be learned by maximizing the conditional log-likelihood of the data (i.e., Eqn. 5). Because there is no analytical solution, we use the BFGS Quasi-Newton for the optimization [13]. The method requires the objective function and its gradients. The partial derivatives of the log-likelihood L with respect to α_i and β_j are given as

$$\frac{\partial L}{\partial \alpha_i} = \sum_m^M \sum_k^K \left(\frac{r_{mk} - P_r}{P_r(1 - P_r)} \sum_{t=1}^n \sigma_\alpha (1 - \sigma_\alpha) \sigma_\beta f_i(q_k, d_t) \right)$$

$$\frac{\partial L}{\partial \beta_j} = \sum_m^M \sum_k^K \left(\frac{r_{mk} - P_r}{P_r(1 - P_r)} \sum_{t=1}^n \sigma_\beta (1 - \sigma_\beta) \sigma_\alpha g_j(e_m, d_t) \right)$$

where P_r, σ_α and σ_β denote the probabilities of Eqn. 6, Eqn. 7, and Eqn. 8, respectively. The main computation of the gradient method is evaluating the log likelihood function and its gradients against parameters. Both of them have computational complexity of $O\big(MKn(N_f + N_g)\big)$. In practice, we only have a small number of relevance judgments for training and thus K is relatively small. In addition, the number of documents associated with each expert and the number of features used are also usually relatively small. Therefore, the training procedure can be efficient.

We can see that both Model 2 and this discriminative model try to aggregate document evidence and document-candidate associations through the bridge of documents, but they are different in how to estimate these two probabilities. In Model 2, the document evidence (i.e., $P(q|d_t)$) is calculated by standard language models and the document-candidate associations (i.e., $P(d_t|e)$) are estimated by a heuristic combination of document-candidate association features. In our proposed discriminative model, both quantities are modeled by logistic functions with arbitrary features and the parameters are automatically determined from training data. From Eqn. 6, we can see that $P_\theta(r = 1|e, q)$ is essentially the arithmetic mean of $P(r = 1|q, d, e)$ with respect to d. Thus we refer the model as the arithmetic mean discriminative (AMD) model.

4.3 An Alternative Discriminative Model with Geometric Mean

It has been shown that in certain cases geometric mean (the product rule) is better than arithmetic mean (the sum rule) in combining evidences [35]. This observation motivates an alternative discriminative model which we refer as the geometric mean discriminative (GMD) model where $P_\theta(r = 1|e, q)$ is modeled by the geometric mean as follows:

$$P(r = 1|e, q) = \frac{1}{Z} \prod_{t=1}^n \Big(P(r_1 = 1|q, d_t) P(r_2 = 1|e, d_t) \Big)^{\frac{1}{n}} \tag{9}$$

where Z is the normalization factor that scales the geometric mean to be a proper probability distribution as follows

$$Z = \sum_{r_1 \in \{0,1\}, r_2 \in \{0,1\}} \prod_{t=1}^n \Big(P(r_1|q, d_t) P(r_2|e, d_t) \Big)^{\frac{1}{n}} \tag{10}$$

Both $P(r_1 = 1|q, d_t)$ and $P(r_2 = 1|e, d_t)$ here take the same form with Eqn. 7 and Eqn. 8. By plugging them and Eqn. 10 into Eqn. 9, we can get

$$P(r = 1|e, q) = \frac{1}{1 + \exp(-E) + \exp(-F) + \exp(-G)} \tag{11}$$

where

$$E = \sum_{i=1}^{N_f} \alpha_i \big(\frac{1}{n} \sum_{t=1}^n f_i(q, d_t) \big), F = \sum_{j=1}^{N_g} \beta_j \big(\frac{1}{n} \sum_{t=1}^n g_j(e, d_t) \big)$$

$$G = \sum_{i=1}^{N_f} \alpha_i \big(\frac{1}{n} \sum_{t=1}^n f_i(q, d_t) \big) + \sum_{j=1}^{N_g} \beta_j \big(\frac{1}{n} \sum_{t=1}^n g_j(e, d_t) \big)$$

We can notice that in Eqn. 11 there are three exponential terms in the denominator, which means that either query-document features $f_i(q, d_t)$ or document-candidate features $g_j(e, d_t)$ alone cannot dominate the final relevance $P(r = 1|e, q)$. The parameters of the model can also be estimated by maximizing the conditional log-likelihood function using BFGS. The GMD model has the same computational complexity with AMD.

4.4 Advantages of Discriminative Models for Expert Search

Some theoretical results show that discriminative models tend to have a lower asymptotic error [24]. Besides the theoretical considerations, we believe there are specific reasons for the domain of expert search that make discriminative models a suitable choice. First of all, the proposed discriminative models can effortlessly incorporate features. As shown in Section 2 and prior research, expert search can benefit from including various types of features. Language modeling approaches often require many modeling assumptions and extra modeling effort to include new features especially when the heterogeneous features are present. Secondly, discriminative models typically make fewer model assumptions than their generative counterparts. For example, many state-of-the-art generative models, including Model 2, the candidate-generation model [14] and the two-stage language model approach [8], assume that the query q and candidate e are independent given the document d, i.e., $p(e|q, d) = p(e|d)$. It requires extra modeling effort for these models to overcome the assumption [4]. In contrast, our proposed discriminative models can easily get around it. For example, $P(r_2 = 1|e, d_t)$ in Eqn. 6 can be replaced by $P(r_2 = 1|e, q, d_t)$ where no independence assumption is made on $P(r_2 = 1|e, q, d_t)$. Thirdly, the discriminative models directly and naturally characterize the notion of relevance. In Model 2 and many other language models, there is no explicit reference to the class variable that denotes whether an expert is relevant or not. We use $P(r = 1|e, q)$ instead of $P(e|q)$ to make it explicit that the relevance of an expert is measured with respect to a query. This explicit notion of relevance can help quantify the extent to which a user's information need is satisfied.

5. EXPERIMENTS

5.1 Data Collections

Our experiments are carried out in the setting of the Expert Search task of the TREC Enterprise tracks from 2005 to 2008. For TREC 2005 and 2006, the document collection was a crawl of the World Wide Web Consortium (W3C) [12, 32]. For TREC 2007 and 2008, a different and more realistic corpus was introduced, which is a crawl of the website of Commonwealth Scientific and Industrial Research Organization (CSIRO). The corpus is known as the CSIRO Enterprise Research Collection (CERC) [1, 7]. Table 1 gives detailed statistics of the collections and query sets. The W3C data is supplemented with a list of 1092 candidate experts represented by their full names and email addresses while the CERC data do not contain a predefined list of candidates. 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. We also use heuristic rules to filter non-personal addresses (e.g. education.act@csiro.au). The total number of candidates extracted is 3,482. In 2005, 50 queries were created based on the working groups in W3C (there were 10 training topics also available in 2005). In 2006, 49 queries were developed by the track participants collectively using the provided list of supporting documents for each candidate. The 50 queries used in 2007 were created with the help of CSIRO's Science Communicators, while the judgments of 77 queries in 2008 were made by participants.

To evaluate the proposed models on W3C, we use the TREC 2006 topics plus the 10 available TREC 2005 training topics for training and test the models on the TREC 2005 topics. Similarly on CERC, we use TREC 2008 topics for training and TREC 2007 topics for testing. Although different years have different ways of topic assessments, we will see in the experiments that the discriminative models can still gain significant improvements from the training data. Our decision of choosing the training and testing configurations is mainly based on the number of relevance judgments available. We need a reasonable amount of training data for the discriminative models and there are relatively more relevance judgments in 2006 for W3C and in 2008 for CERC. Because the two test collections have very different characteristics, we do not evaluate the models across the corpora. To obtain a balanced training set, we randomly select the same number of negative instances with the number of positive instances for each training query, by following the undersampling method in [23]. To acquire negative instances for the queries without non-relevance judgments (i.e., 10 TREC 2005 training topics), we use the Base method introduced in Section 6.1 to identify a list of unjudged/irrelevant experts for each query. Evaluation measures are mean average precision (MAP), R-precision (R-Prec), mean reciprocal rank (MRR), and precision@5 (p@5) and precision@10 (p@10).

Table 1: Statistics of the W3C and CERC testbeds

	W3C	CERC
# Documents	331,037	370,715
# People	1,092	3,482
Avg. Doc Length in Token	983.4	354.8
Avg. # Rel Experts/Topic (TREC Year)	51.5 (2006) 30.2 (2005)	10.4 (2008) 3.0 (2007)
Training Queries	2006 (49) 2005 (10)	2008 (77)
Testing Queries	2005 (50)	2007 (50)

5.2 Research Questions

An extensive set of experiments were designed to address the following questions of the proposed research:

- Can the discriminative trained model perform better than its generative counterpart when the same set of features are available for use? (Section 6.1)
- Can integration of additional features into the discriminative model improve the performance? (Section 6.1)
- What features are likely more important in terms of the relative values of the learned weights in the discriminative model? (Section 6.1)
- What is the effect of only retrieving a subset of documents on the proposed model? (Section 6.2)
- How robust is the proposed discriminative model with respect to the underlying document retrieval methods? (Section 6.3)
- How robust is the proposed discriminative learning framework with respect to specific discriminative models? (Section 6.4)

In all the sections except Section 6.4, we only use the arithmetic mean discriminative (AMD) model to assess the discriminative learning approach, since we care less about the difference between discriminative models than about the difference between generative and discriminative models.

5.3 Experimental Setup

In all our experiments, we have done minimal preprocessing in which both queries and documents are stemmed using Krovetz stemmer. We only use the "title" or "query" fields in the topics without using extra information (e.g., "narrative"). No query expansion nor external resource is utilized. As shown in Section 4, each query-expert pair is characterized by two feature vectors, i.e., document evidence $f_i(q, d_t)$ and document-candidate associations $g_j(e, d_t)$. Table 2 summarizes the features used in the discriminative models.

These features include the score from the standard document language model (f_1), document features ($f_2 - f_5$), external document structure features ($f_6 - f_9$), basic association features ($g_1 - g_5$), internal document structure features ($g_6 - g_9$), and proximity features ($g_{10} - g_{13}$). Here the external document structure features are the boolean variables to represent whether a document (in W3C) comes from specific types of documents (e.g., $f_8 = 1$ means the document is either from "www" or "esw"). The evaluations on W3C use all the features, while the features $f_6 - f_9$ and $g_6 - g_9$ are not applied to CERC, as the CERC dataset does not

Table 2: Features used in the discriminative models. "B" denotes the feature takes boolean values and "N" represents numerical values

Feature	Description	Type	References
f_1	LM	N	[37]
f_2	PageRank	N	[38]
f_3	URL length	N	[38]
f_4	Anchor text	N	[38]
f_5	Title	N	[38]
f_6	From lists	B	[12]
f_7	From people	B	[12]
f_8	From www+esw	B	[12]
f_9	From other+dev	B	[12]
g_1	Exact name match	B	[3]
g_2	Name match	B	[3]
g_3	Last name match	B	[3]
g_4	Email match	B	[3]
g_5	LM score	N	[6]
g_6	EMAIL_FROM	B	[5]
g_7	EMAIL_TO	B	[5]
g_8	EMAIL_CC	B	[5]
g_9	EMAIL_CONTENT	B	[5]
$g_{10} \sim g_{13}$	Proximity	B	-

Table 3: Experimental configurations

Base	Balog et al's Model 2 (candidate-centric) with 4 association features (i.e., $g_1 - g_4$) [3]
R1	Discriminative model with 4 association features ($g_1 - g_4$) and LM document evidence feature (f_1)
R2	Discriminative model with full document evidence features and 4 association features ($g_1 - g_4$)
R3	Discriminative model with full association features and one document evidence feature (f_1)
R4	Discriminative model with full document evidence features and full association features

Table 4: Comparison of the discriminative model (AMD) with the Base mehod on W3C and CERC. Best results on each collection are highlighted. The †symbol indicates statistical significance at 0.95 confidence interval against Base

	MAP	R-Prec	MRR	P5	P10
W3C					
Base	0.1909	0.2445	0.5081	0.3760	0.3120
R1	0.2001	0.2552	0.5300	0.3820	0.3310
R2	0.2282†	0.2764	0.5624†	0.3960	0.3370
R3	0.2412†	0.2904†	**0.6232**†	0.4020	0.3560†
R4	**0.2598**†	**0.3035**†	0.6196†	**0.4130**†	**0.3680**†
CERC					
Base	0.4039	0.3514	0.5389	0.2240	0.1540
R1	0.4123	0.3569	0.5593	0.2280	0.1540
R2	0.4453†	0.3854†	0.5924†	0.2390	0.1650
R3	0.4569†	0.3879†	0.5886†	**0.2610**†	0.1660
R4	**0.4604**†	**0.3938**†	**0.6143†**	0.2520†	**0.1770**†

contain explicit document types nor many emails with internal structure information useful for expert search [38]. The f_1 feature is the document retrieval score by LM using the topic as the query. The smoothing method of LM is Jelinek-Mercer with the parameter $\lambda = 0.5$ (we use the same smoothing for other LMs). The g_5 feature is the retrieval score by LM using the candidate identifier as the query [6]. The "Proximity" features ($g_6 - g_9$) are the boolean variables indicating whether the candidate identifier co-occurs with the query term in a window with various sizes. We use 20, 50, 100 and 250 as the window sizes (in number of words), approximated to the sizes of sentence, passage, paragraph and section, respectively. The details about these features can be found in the corresponding reference. To normalize the features, we use query-based normalization for each feature as suggested in [19].

Many of these features have been shown useful for expert search. Because of the generative nature of language models, it is difficult for them to incorporate such heterogeneous features in a unified modeling framework, but discriminative models can effortless include all the features and many more. Since the focus of this study is on the probabilistic models rather than feature engineering, we do not intend to choose a complete set of features, but they are one of the most comprehensive and diverse feature sets in a single work among the existing expert search research.

6. RESULTS

6.1 Discriminative Model vs. Model 2

In this section, we compare the proposed discriminative model with its generative counterpart: Model 2. The proposed model is evaluated on four different feature configurations, which are presented in Table 3. The Base method is the implementation of Model 2 by following [3], which includes 4 types of document-candidate associations. The R1 configuration uses these 4 association features plus f_1 as document evidence. Thus, the identical information is available for R1 and Base to use. The weights in Base are set by following the choice of the best run in [3]. R4 is the configuration with full applicable features for the discriminative model (the R4 configuration is the default setting in all the experiments except explicitly noted). Table 4 contains the evaluation results on the two test collections. We can see that the discriminative model consistently performs better than Base across all the feature configurations on all measures. With the full set of features (i.e., R4 vs Base), all the differences are statistically significant by two-tailed Student's t-test at 0.95 confidence level. In R1 vs Base, although their differences are not significant, the discriminative model outperforms the Base method on all the evaluation metrics.

Since all the features are normalized, the weight associated with each feature can reflect the importance of the feature in some degree. Table 5 reports the top 3 features with the largest weights in f_i and g_i respectively in the learned AMD model. These features are ordered alphabetically in the table since their weights are not very distinct from each other. We find that the features listed for the two testbeds are generally different with the exception of f_1 and f_2, showing the importance of these two features across the corpora. An interesting observation is that the g_8 feature when used on W3C has a large weight among all the document-candidate association features. This is intuitive in the sense that the person who is in the email cc field is likely an authoritative of the topics of the email, which is also consistent with what was reported in [5]. Another observation is that the "Proximity" features have large weights for both testbeds (i.e., g_{13}

Table 5: The top 3 features with the largest weights in AMD (R4) learned from training data

	Doc evidence	Doc-candidate associations
W3C	f_1, f_2, f_6	g_1, g_8, g_{13}
CERC	f_1, f_2, f_5	g_4, g_5, g_{11}

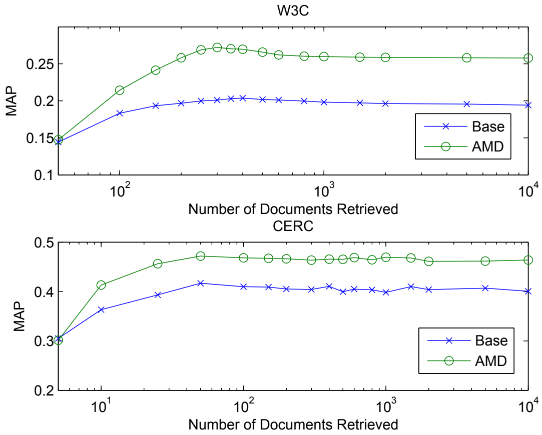

Figure 1: Impact of varying the number of documents retrieved (M) on the discriminative model. Top: impact on W3C; Bottom: impact on CERC.

for W3C and g_{11} for CERC), but with different window sizes: i.e., larger size on W3C. This may come from the fact that these two collections have very different average document lengths.

6.2 The Effect of the Size of Retrieved Documents

Similar to Model 2, the learned discriminative model can be efficiently used on top of an existing document search engine as follows: 1) Perform a standard document retrieval run using the topic as a query and retrieve the top m documents; 2) For each candidate associated with the relevant documents, calculate the probability of relevance using Eqn. 6 on these m documents. In this section, we aim to investigate the effect of the size of documents retrieved on the performance of the discriminative model. We use LM as the document retrieval run. Figure 1 shows the MAP results by varying M on the two test collections. Note that the scales on the x-axis and y-axis differ per plot. From the figure, we can see that as M increases, the discriminative model has a similar trend with the baseline: increasing, achieving a maximum, and then flattening. On W3C, the MAP value tops after 300 documents retrieved, fewer than what the baseline needs (i.e., 400). For CERC, both models need around 50 documents for best performance. Therefore, using a subset of documents could speed up the process of expert search as the best performers use much less documents than the whole set of relevant documents. At the same time, the retrieval performance can be improved although their differences are not found statistically significant.

6.3 Experiments by Using Different Document Retrieval Methods

As shown in Section 6.1 as well as in prior work, the document retrieval score f_1 is an important feature to show document evidence for expert search. In this experiment, we assess the extent to which the performance of the discriminative model is affected by the choice of the underlying document retrieval model. Besides LM, another two different document retrieval methods are used (i.e., BM25 [28] and Indri [33]). Specifically, the f_1 feature is replaced by these two retrieval scores respectively in the R4 configuration. Table 6 shows the MAP results of the proposed model across the three retrieval models. From the table, we can see that the results are quite similar and they are all significantly better than the baseline. This indicates that the discriminative model is robust to the underlying document retrieval method.

Table 6: Evaluation of AMD with different document retrieval methods on W3C and CERC

	MAP	R-Prec	MRR	P5	P10
W3C					
LM	0.2598	0.3035	0.6196	0.4130	0.3680
BM25	0.2658	0.3141	0.6238	0.4060	0.3700
Indri	0.2562	0.3066	0.6149	0.4090	0.3640
CERC					
LM	0.4604	0.3938	0.6143	0.2520	0.1770
BM25	0.4551	0.3895	0.5877	0.2470	0.1740
Indri	0.4667	0.4086	0.6000	0.2550	0.1780

Table 7: Comparison of the geometric mean discriminative model with Base and AMD (R4) on W3C and CERC. The †symbol indicates statistical significance at 0.95 confidence interval for GMD against Base

	MAP	R-Prec	MRR	P5	P10
W3C					
Base	0.1909	0.2445	0.5081	0.3760	0.3120
AMD	0.2598	0.3035	0.6196	0.4130	0.3680
GMD	0.2512†	0.3010†	0.6266†	0.4110†	0.3640†
CERC					
Base	0.4039	0.3514	0.5389	0.2240	0.1540
AMD	0.4604	0.3938	0.6143	0.2520	0.1770
GMD	0.4669†	0.4030†	0.6274†	0.2500†	0.1790†

6.4 The Alternative Discriminative Model vs. Base and AMD

In this section, we conduct the experiment to evaluate the alternative discriminative model (GMD). The aim is to investigate the robustness of the proposed discriminative framework with respect to the choice of specific discriminative models derived from the framework. Table 7 contains the results. From the table, we can see that all the results achieved by GMD significantly outperform the baseline. Furthermore, these results are quite similar with those achieved by the AMD (R4) model. In particular, the GMD model is generally better than AMD on CERC and worse on W3C, but the differences between GMD and AMD are not statistically significant. These results demonstrate that the proposed discriminative framework generates accurate and robust results with both types of discriminative models.

7. CONCLUSIONS AND FUTURE WORK

In this work, we propose a discriminative learning framework and derive specific models for expert search. The main advantage of the proposed approaches is their ability to integrate a variety of document evidence and document-candidate association features. The evaluations on two TREC Enterprise track testbeds have shown the effectiveness and robustness of the proposed framework.

There are several possibilities to extend the research in this paper. We chose "out-of-order" training in the experiments because more training data are available in 2006 and 2008. It would be interesting to perform the "in-order" experiments (i.e., training on 2005 or 2007), which would allow fair comparisons with the TREC submitted runs. The relevance judgments in 2005 and 2007 seem also more likely to be obtained in a real enterprise. In fact, lack of training data hinders the applicability of many discriminative models. On the other hand, generative models may be able to effectively utilize abundant unlabeled data. It is desirable to develop a hybrid of discriminative and generative models to obtain the best of both for expert search. In addition, in certain scenarios, pairwise comparisons between experts might be more easily collectible than the pointwise judgment for each expert. We will explore to extend the proposed discriminative learning framework to handle this type of training data.

8. ACKNOWLEDGMENTS

We thank the anonymous reviewers for many valuable comments. This research was partially supported by a grant from the Indiana Economic Development Company, the NSF research grant IIS-0749462, and a grant from Purdue University. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] P. Bailey, N. Craswell, A. De Vries, and I. Soboroff. Overview of the trec-2007 enterprise track. In *TREC-15*, 2007.

[2] K. Balog. Non-local evidence for expert finding. In *CIKM*, 2008.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR*, 2006.

[4] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19, 2009.

[5] K. Balog and M. de Rijke. Finding experts and their details in e-mail corpora. In *WWW*, page 1036. ACM, 2006.

[6] K. Balog and M. De Rijke. Associating people and documents. In *ECIR*, 2008.

[7] K. Balog, I. Soboroff, P. Thomas, N. Craswell, A. de Vries, and P. Bailey. Overview of the trec-2008 enterprise track. In *TREC-16*, 2008.

[8] Y. Cao, J. Liu, S. Bao, and H. Li. Research on expert search at enterprise track of TREC 2005. In *TREC-13*, 2005.

[9] P. Carlile. Working knowledge: how organizations manage what they know. *Human Resource Planning*, 21(4):58–60, 1998.

[10] H. Chen, H. Shen, J. Xiong, S. Tan, and X. Cheng. Social network structure behind the mailing lists: Ict-iiis at trec 2006 expert finding track. In *TREC-14*, 2006.

[11] W. Cooper. Exploiting the maximum entropy principle to increase retrieval effectiveness. *JASIST*, 34(1):31–39.

[12] N. Craswell, A. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *TREC-13*, 2005.

[13] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial Mathematics, 1996.

[14] H. Fang and C. Zhai. Probabilistic models for expert finding. In *ECIR*, 2007.

[15] Y. Fang, L. Si, and A. Mathur. Ranking experts with discriminative probabilistic models. In *SIGIR Workshop on Learning to Rank for Information Retrieval*, 2009.

[16] Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang, and S. Ma. THUIR at TREC 2005: Enterprise track. In *TREC-14*, 2006.

[17] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243, 1992.

[18] T. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[19] T. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR Workshop on Learning to Rank for Information Retrieval*, 2007.

[20] C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *ECIR*, 2008.

[21] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM*, 2006.

[22] D. Metzler and W. Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[23] R. Nallapati. Discriminative models for information retrieval. In *SIGIR*, 2004.

[24] A. Ng and M. Jordan. On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. *NIPS*, 2002.

[25] D. Petkova and W. Croft. Proximity-based document representation for named entity retrieval. In *CIKM*, 2007.

[26] S. Robertson. The probability ranking principle in IR. *Journal of documentation*, 33(4):294–304, 1977.

[27] S. Robertson and K. Jones. Relevance weighting of search terms. *JASIST*, 27(3):129–146, 1976.

[28] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-4. In *TREC-4*, 1996.

[29] P. Serdyukov and D. Hiemstra. Being omnipresent to be almighty: The importance of the global web evidence for organizational expert finding. In *SIGIR Workshop on Future Challenges in Expertise Retrieval*, 2008.

[30] P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR*, 2008.

[31] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM*, 2008.

[32] I. Soboroff, A. de Vries, and N. Craswell. Overview of the trec-2006 enterprise track. In *TREC-14*, 2006.

[33] T. Strohman, D. Metzler, H. Turtle, and W. Croft. Indri: A language model-based search engine for complex queries. In *International Conference on Intelligence Analysis*, 2004.

[34] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *SIGKDD*, 2008.

[35] D. Tax, M. Van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern recognition*, 33(9):1475–1485, 2000.

[36] D. Yimam-Seid and A. Kobsa. Expert finding systems for organizations. *Sharing Expertise: Beyond Knowledge Management*, 2003.

[37] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS*, 22(2):214, 2004.

[38] J. Zhu, X. Huang, D. Song, and S. Ruger. Integrating multiple document features in language models for expert finding. *Knowledge and Information Systems*, pages 1–26.