Scientific Contribution Learning for Expertise Ranking

ABSTRACT

Expert finding in bibliographic networks has received increased interests in recent years. This task concerns with finding relevant researchers for a given topic. Motivated by the observation that rarely do all coauthors contribute to a paper equally, in this paper, we propose a discriminative method to realize leading authors contributing in a scientific publication. Specifically, we cast the problem of expert finding in the bibliographic network to find leading experts in a research group, which is easier to solve. According to some observations, we recognize three feature groups that can discriminate relevant and irrelevant experts. Experimental results show that the proposed method significantly improves the state-of-the-art model of expert finding in terms of all common Information Retrieval evaluation metrics.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*

General Terms

Algorithms, Experimentation

Keywords

Expertise retrieval, Contribution learning, Learning to Rank, Discriminative models

1. INTRODUCTION

As the large portion of web provides information for various kinds of real-world objects (i.e. entities), more search engines provide object level search result. Typical objects are products, peoples, papers, organizations, and etc.. As one of the noteworthy resources in the world, searching the human expertise has recently attracted much attention in Information Retrieval (IR) community. Considering human experts as the web objects, expert finding is one of the challenging types of object level search, which concerns itself with ranking people who are knowledgeable in a given topic.

Initial approaches for expert finding have been proposed to identify experts in simple environments such as organizations [2], universities [3] and etc.. While these approaches are quite effective in these simple domains, they are not appropriate for complicated environments such as bibliographic networks [10], hierarchical organizations [14] and expert's social networks [23]. Considering associated documents of each expert as the main evidence of his or her expertise, these approaches are not able to take into account other valuable expertise evidences in such complicated domains. These evidences can be social interactions between peoples, temporal behaviors of them, the document quality indicators, and etc., which are usually independent of the content of the documents.

Identification of knowledgeable persons in a specific academic field could be of great value in many applications such as recognizing qualified experts to supervise new researchers, assigning a paper to reviewers [18], and expert team formation [15]. As a complicated search domain, bibliographic networks contain various types of documents (e.g. conference proceedings, journal articles), experts (e.g. students and supervisors) and relationships between them (e.g. co-author and citation relationships) that can be used to infer the expertise of the academic persons. Finding experts in a bibliographic network is a challenging task because of the following reasons:

1. In contrast with simple domains, there exist a huge number of documents and expert candidates in a bibliographic network, making it difficult to find the distinguished experts among numerous expert candidates.

2. Besides the content of associated documents of a person in these networks, it is necessary to consider some other important factors such as the quality of documents, social interactions and temporal behavior of authors for expert ranking. As a result, identification of these effective features is an important step toward building a high-quality search engine for bibliographic networks.

3. While scientific publications are usually a product of cooperation among members of a research group, it is not obvious how to determine the contribution of each author in a given paper. Indeed, precisely estimating the contribution of each author in a multi-author publication is beneficial to recognize leading experts and this is one of the main challenges of expert finding for bibliographic networks.

Many approaches have been proposed and shown to be effective for expert finding. For example, document centric models [2, 10] share the underlying claiming that the relevance of the textual context of a person adds up to the evidence of his or her expertness. Although these methods are so beneficial for expert finding, they only consider the equivalent share for each expert candidate mentioned in a scientific publication. Generally, the authors of a paper have different expertise in the domain of that paper. For instance, while the supervisor of a PhD project has a broader view of the research problem, other co-authors might be more involved in the detail of the project. As another example, contributing in an industrial project, the project managers have more expertise in the problem domain than the other team members. While recognizing the contribution of each author in a scientific publication has been noticed in many researches [1, 20], it is not considered in expert finding approaches.
In this paper, we propose a learning algorithm to determine the contribution of each author in a multi-author publication that significantly improves the ranking of experts. We examine the impact of various features to estimate the contribution of authors in a bibliographic network. Specifically, we consider three feature groups; namely, structural, temporal, and activity-based features. We found that all of these feature groups are beneficial to determine the scientific contribution of authors. We compare our proposed models with the baseline and the state-of-the-art algorithms for expert finding in bibliographic networks.
The rest of this paper is organized as follows. In Section 2 we review some related work on expertise search. Section 3 is devoted to a description of the background and preliminaries, and in Section 4 we detail our models of scientific contribution mining. In Section 5, we define the experimental setup and report the experimental results. Finally, we present the conclusions and future work in Section 6.

2. RELATED WORK

The inclusion of the expert finding task in TREC Enterprise has attracted a significant amount of attention from 2005 to 2008 [8, 22, 6]. The main approaches of expert finding can be categorized into two groups: profile centric and document centric approaches, were originally proposed in [2]. While profile centric method (i.e. *Model 1* in [2]) directly models the knowledge of an expert from associated documents, document centric method (i.e. *Model 2* in [2]) first locates documents on the topic and then finds the associated experts. Document centric method is generally more effective and easier to implement than profile centric method [5]. Therefore, most recent methods for expert finding are basically the extension of the document centric method.
The problem of expert finding in bibliographic networks has been introduced in [10]. In this research, the goal is to retrieve experts in specific academic domains based on their publications in the DBLP [1] bibliographic network. More recently, Deng et al. [11] proposed a new smoothing method based on the community context of each author and also a community-sensitive AuthorRank method for co-authorship networks. They observe that the community provides valuable and distinctive information along with the documents and the experts. They also studied the expertise ranking problem through modeling and exploiting heterogeneous network together with the textual content information [9].
In all above mentioned approaches [9, 10, 11], if a document is associated with more than one expert candidate, then all of these candidates will get an equal score from that document. In contrast, some approaches consider a non-equal expertise scores for expert candidates occurred in a single document. For example, Serdyukov and Hiemstra [21] proposed an expert centric language model which assumes that the terms in a document are generated by those persons who are mentioned in it. For documents with more than one associated candidate, they proposed an EM-algorithm to refine the specific language model of each expert candidate. As another approach, Petkova and Croft [19] considered the proximity of query terms to expert's name occurrences and give more score to the expert candidates whom their names are occurred near to the query terms.
Similar to these approaches, our method assigns non-equal scores for expert candidates associated with a document. While the proposed method in [21] is an unsupervised approach and basically a profile centric method, our model is based on the document centric approach which utilizes supervised learning to determine the author's contributions. Besides the simplicity and the performance of the document centric methods in comparison with the profile centric method, our proposed model is able to employ various features to learn author's contribution. On the other hand, in bibliographic networks, the names of authors are usually mentioned in the same position of documents (just after title), so, the proximity based method proposed in [19] cannot be useful to determine the contribution of each author in a document.
Balog and Rijke [4] proposed three methods (i.e. boolean, frequency based and semantic relatedness) to estimate the strength of association between a candidate expert and a document. The boolean method simply assigns zero/one association weights according to occurrence of an expert's name in a document; however, the frequency method assigns weights based on the occurrence number of an expert's name in a document. To assess the frequency based method, they introduce semantic relatedness of the document and the person. While they find that frequencies capture the semantics of person-document associations very well, they prefer to utilize frequency based methods in their document based models. These methods have different behaviors on TREC test collections; nevertheless, they assigns equal weights to the candidate-document association in bibliographic networks. Indeed, the boolean weighting method is used in previous research of expert finding in bibliographic networks [9, 10, 11].
While academic search engines calculate rankings of experts without regard to authors' contribution, some previous researches [20, 1] suggested that authors of a document should not necessarily get equal contribution. For example, in [20], Sekercioglu suggested that the contribution of kth coauthor should be 1/k as much as the first author. In [1], Abbas introduced weighted h-indices that take into account the weighted contributions of individual authors in multi authored papers. While these approaches estimate the contribution of each expert based on their names' order, we used extensive set of features (i.e. structural, temporal, and activity-based) to learn the contribution of authors in research publications.
As another related line of research, Fang et al. [12] proposed

[1] www.informatik.uni-trier.de/ ley/db

a discriminative learning framework for expert search. Their model is able to integrate a variety of document's evidence and document association features. More recently, Macdonald and Ounis [17] proposed an approach to learn rankings for aggregate search tasks such as expert search. However, these methods did not take into account unequal contribution assignment for authors of multi-author documents.

3. BACKGROUND AND PRELIMINARIES

As the most effective approach for expert finding, document-centric model [2, 10] estimates the expertise of an expert candidate by summing the relevance of the associated documents. In this section, two main baseline approaches for expert ranking in bibliographic networks are described. The first method is *Model 2* proposed by Balog et al. [2] and the second one is the weighted language model proposed by Deng et al. [10] which is the state-of-the-art model of document centric expert finding.

3.1 Baseline Document Centric Model (ULM)

In this model, for a given query q, the relevance probability of each expert is determined by the following equation:

$$p(e|q) = \frac{p(q|e)p(e)}{p(q)}, \tag{1}$$

in which, $p(e|q)$ represents the score of a candidate e for the given query q, $p(q|e)$ is the probability of a query q given the expert candidate e, $p(e)$ is the prior probability of expert candidate e, and $p(q)$ is the prior probability of a given query q. As $p(q)$ is a constant in expert ranking, it can be ignored from above equation. Therefore, relevance probability of each experts can be estimated by the probability of a query given expert candidate ($p(q|e)$), weighted by the prior probability that expert candidate e is an expert ($p(e)$):

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

In the document centric methods [2, 12], the probability of a query given expert candidate can be estimated by the following equation:

$$p(q|e) = \sum_{d \in D_e} p(d|e)p(q|d,e). \tag{3}$$

In this equation, following the assumption of Balog et al. [2], we assume that expert candidate e is conditionally independent of the query q given a document d (i.e. $p(q|d,e) \approx p(q|d)$). As a result, by substituting $p(q|e)$ in Equation (2), $p(e|q)$ can be estimated as follows:

$$\begin{aligned} p(e|q) &= \sum_{d \in D_e} p(e)p(q|d)p(d|e) \\ &= \sum_{d \in D_e} p(e)p(q|d)\frac{p(e|d)p(d)}{p(e)} \\ &= \sum_{d \in D_e} p(d)p(q|d)p(e|d), \end{aligned} \tag{4}$$

where D_e indicates the subset of documents associated with the expert candidate e, $p(d)$ denotes the prior relevance probability of document d, $p(q|d)$ is the probability of a query q given the document d, and $p(e|d)$ is the probability of the association between the document d and the expert candidate e. The probability $p(e|d)$ provides a ranking of candidates associated with a given document d, based on their contribution made to d [4]. According to equation (4), in order to rank candidate experts, we should estimate three probabilities, namely, prior probability of retrieval for document d (i.e. $p(d)$), relevance probability of document d (i.e. $p(q|d)$), and association probability of document d and expert candidate e (i.e. $p(e|d)$).

The relevance probability of a document d to the query q (i.e. $p(q|d)$ in equation (4)) can be estimated by the document language model θ_d of document d:

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)^{n(t,q)}, \tag{5}$$

where $p(t|\theta_d)$ is the probability of a term t given the document model θ_d, and $n(t,q)$ is the number of times that term t occurs in query q. In order to overcome zero probabilities, we use the JM-smoothing [24], so the probability $p(t|\theta_d)$ is estimated as $p(t|\theta_d) = (1-\lambda)p(t|d) + \lambda p(t)$, where $p(t|d)$ is the maximum likelihood estimation of the occurrence of term t in the document d, and $p(t)$ is the occurrence probability of the term t in the document repository. In our experiments, we follow Balog et al. [2] in setting the smoothing parameter $\lambda = 0.5$. Balog et al. [2] assume the document prior probability $p(d)$ is uniform, so it does not affect the ranking of experts. This model is based on the assumption that expert candidate e is knowledgeable about the topics of document d if an expert candidate e is an author of it. For a multi-author document d, it assumed that each author has the same level of knowledge about the topics described in the document and therefore the association probability of document d and expert candidate e is estimated as follows:

$$p(e|d) = \begin{cases} \frac{1}{n_d} & \text{e is an author of d} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where n_d is the number of authors of document d. To sum up, by substituting equations 5 and 6 in equation 4, the final estimation of the Balog's language model is obtained by the following equation.

$$p(e|q) \cong \sum_{d \in D_e} \left(\prod_{t \in q} p(t|\theta_d)^{n(t,q)}\right)\frac{1}{n_d} \tag{7}$$

In the rest of the paper, we refer to this approach as the uniform language model (ULM).

3.2 Weighted Language Model (WLM)

The method given in Section 3.1 calculates the relevance probability between the query and expert candidate, but it ignores the prior relevance probabilities of the documents. If we assume that two documents d_1 and d_2 have similar contents, then the query likelihoods are almost the same (i.e. $p(q|d_1) = p(q|d_2)$), therefore, the ULM assigns the same score to the authors of these documents. However, if these documents have different importance, we would prefer to rank the authors of the more important document higher than the less important one. This is the main idea of the weighted language model proposed by Deng et al. [10]. They introduced a weight factor w_d to denote the importance of a document, which, theoretically, can be interpreted as being proportional to the document prior probability $p(d)$. The weight factor is estimated using the citation number and is transformed by the natural logarithm.

$$w_d = \ln(e + c_d)$$

where c_d ($c_d \geq 0$) is the citation number of the document d and constant e is used to guaranty that the weight factor not to be less than 1. The final estimation of the weighted language model is:

$$p(e|q) \cong \sum_{d \in D_e} w_d (\prod_{t \in q} p(t|\theta_d)^{n(t,q)}) \frac{1}{n_d}$$

In the rest of the paper, we refer to this approach as the Weighted language model (WLM).

4. LEARNING SCIENTIFIC CONTRIBUTION

The document centric method introduced in section 3 is an effective method for finding experts in simple environments that each document usually associated to a single author. In these environments, the boolean weighting model [4] can efficiently estimate the document expert association probability (i.e. $p(e|d)$ in equation (4)) like follows:

$$p(e|d) = \begin{cases} 1 & \text{e is an author of d} \\ 0 & \text{otherwise} \end{cases}$$

Moreover, because the name of each author of a paper is occurred only once in that paper, the frequency based association estimation model [4] has the similar behavior with boolean method in bibliographic networks. In contrast with expert finding in above mentioned environments, a substantial fraction of documents in bibliographic networks is associated to more than a single author. For example, in DBLP bibliographic network, more than 76 percent of documents (1,254,058 of 1,632,442 papers)[2] have more than one author. So, the expert-document association probability (i.e. $p(e|d)$) becomes an effective component to estimate expert relevancy. Surprisingly, previous methods (e.g. [10, 11, 9]) for expert finding in bibliographic networks, simply ignore the effect of this component and assign the same score to all authors of a multi-author document.

Indeed, rarely do all coauthors contribute to a paper equally. While this fact is considered in some researches (e.g. [1, 20]), precisely estimating the contribution of each author can be beneficial for expert finding in bibliographic networks. Suppose a multi-author document d written by two authors a_1 and a_2, here, the question is which author has more contribution in generating the document d? Which author should be ranked higher for a query relevant to document d? Obviously, we would prefer to rank the author with more contribution higher than the rest authors of the document. To the best of our knowledge, current approaches for expert finding in bibliographic network do not take this factor into account. We observe that some suitable features can help significantly to estimate the contribution of authors in research publications. In the rest of this section, first we introduce some effective feature groups, and then propose our method for learning contribution of each author in multi-author documents.

4.1 Feature Groups

Intuitively, various features can be beneficial to determine the contribution of authors in a document. For example, research longevity of an author in field of the document, an expert's authority in a specific research community, and the research activity (e.g. diversity, quality of research, as well as semantic relatedness of the document and the expert) of an author are some effective factors in recognizing distinguished expert in research communities.

[2]The dataset is gathered from the Arnetminer website.

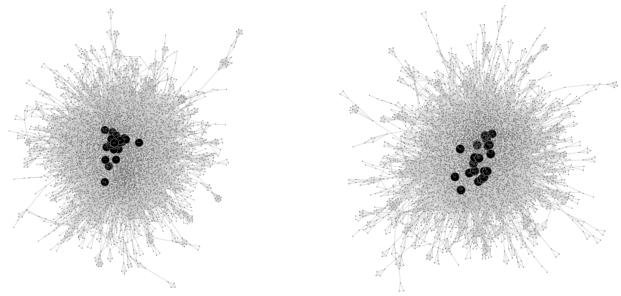

Figure 1: The co-authorship networks of top retrieved authors for Information Retrieval (left figure) and Face Recognition (right figure) topics.

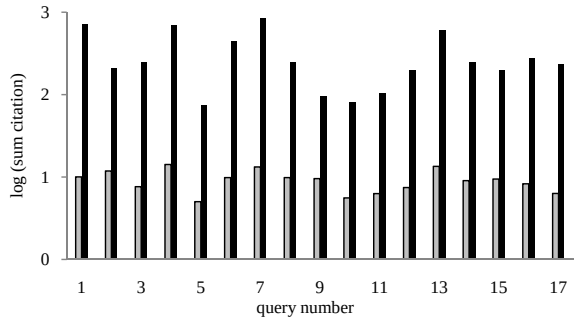

Figure 2: Average of sum of citations for relevant experts and non-relevant authors in logarithm scale.

Figure 1 indicates the co-author network of top retrieved expert candidates using the ULM (introduced in Section 3.1) for two research topics (i.e. Information Retrieval and Face Recognition), in which relevant experts and non-relevant authors are illustrated with black and gray colors, respectively. According to this figure, it is obvious that relevant experts tend to be placed in the center of the co-author network.

As another example, to quantify the research quality of authors, Figure 2 illustrates the average of the sum of citations for relevant experts (black columns) and non-relevant authors (gray columns) retrieved by the ULM. It is obvious that the relevant experts obtain a greater sum of citations than the non-relevant authors.

Intuitively, the research longevity of an expert candidate can be a useful evidence to infer his or her expertise. As indicated in Figure 3, in most of the research topics, the research longevity of relevant experts (black columns) is more than non-relevant authors (gray columns).

Note that these authors (relevant experts and non-relevant authors) are retrieved by the ULM, and this approach is not able to discriminate these authors, because it ignores the effect of the mentioned feature groups. According to above observations, we define three feature groups to discriminate relevant experts from non-relevant authors. These feature groups include: structural, temporal, and activity-based features. Although, these groups of features may have positive correlation with each other, but each group is conceptually

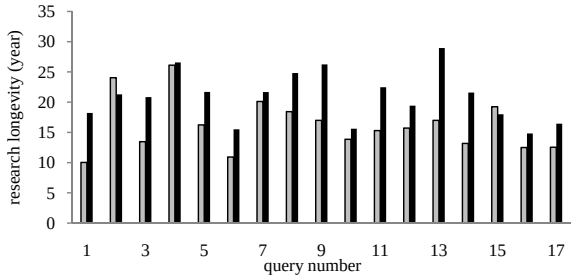

Figure 3: Average of research longevity of relevant experts and non-relevant authors.

distinct from other groups.

4.2 Learning Model

In this section, we propose two discriminative methods to predict the contribution of each author in a multi-author document. The predicted value for the contribution of an author then can be plugged into the equation (4) to estimate the relevancy of each expert on a topic. For a given topic, we cast the contribution estimation problem into a classification problem that treats the relevant experts in a multi-author document as positive data, and non-relevant authors as negative data. Formally, we use a contribution variable $c \in \{0, 1\}$ to denote how much author e of document d has contribution on generating it. Specifically, the probability of $P_\theta(c = 1|e, d)$ can be used as the contribution estimator of author e for paper d, where θ is the unknown parameters that should be learned using training data.

In order to train our model, for a given topic t, we divide the set of retrieved documents by ULM into two distinct subsets. The first subset (i.e. D_{1t}) includes documents which have been written by at least one relevant expert to topic t (an author which is determined as a relevant expert for topic t in the test collection) and the second subset (i.e. D_{2t}) includes all documents that all of their authors are non-relevant to the topic t. By definition of D_{1t} and D_{2t}, we have $D_t = D_{1t} \cup D_{2t}$ where D_t is the set of documents retrieved by the ULM for the topic t.

4.2.1 Contribution learning Model A

While there is no obvious evidence for the relative contribution of the paper's authors in the set D_{2t}, we can assume that relevant experts have more contributions in comparison with non-relevant authors in the set D_{1t},. Therefore, in our first attempt for building a learning model, we only use the documents of D_{1t}. For each document $d \in D_{1t}$, we produce the training set as $T_1 = \{(e, d, e_t)|d \in D_{1t}, e \in A(d), e_t \in \{0, 1\}\}$, where $A(d)$ denotes authors of document d, e indicates an author of it, and e_t is the label of author e in the expert finding test collection for topic t. Specifically, $e_t = 1$ if author e is relevant expert for topic t.

The members of set T_1 can be divided into positive and negative instances. Given the relevance judgment e_t for each expert e on a topic t, the likelihood L of training data is as follows. We assume that the relevance judgments e_t are generated independently.

$$L = \prod_{m=1}^{|T_1|} \prod_{e \in A(d_m)} P_\theta(c = 1|d_m, e)^{e_t} P_\theta(c = 0|d_m, e)^{1-e_t},$$

where $A(d_m)$ denotes the authors of document d_m. We model $P_\theta(c = 1|e, d_m)$ by logistic functions on a linear combination of features. The unknown parameters θ can then be estimated by maximizing the following log likelihood function.

$$\theta^* = \text{argmax}_\theta \sum_{m=1}^{|T_1|} \sum_{e \in A(d_m)} (e_t \log P_\theta(c = 1|d_m, e) + (1 - e_t) \log P_\theta(c = 0|d_m, e)) \quad (8)$$

The estimated parameters can then be plugged back in $P_\theta(c = 1|e, d_m)$.

4.2.2 Contribution learning Model B

In the first learning model, we used only the papers in the set D_{1t} and ignored the papers in the set D_{2t}. In order to avoid the bias and more precisely predict the contribution of each author, we can also use the papers in set D_{2t}. As mentioned before, for papers in set D_{2t} we do not have any evidence that confirms the relative contribution of authors in a multi-author document. So, following the idea of Deng et al. [10], for each document $d \in D_{2t}$, we assume the same contribution for all authors. In this case, the predicated contribution of an author in a given paper can be estimated by the following equation:

$$p(e|d) = P_{\theta_2}(y = 1|d) P_{\theta_1}(c = 1|d, e, y = 1) + (1 - P_{\theta_2}(y = 1|d)) \frac{1}{|A(d)|}, \quad (9)$$

where $P_{\theta_2}(y = 1|d)$ indicates the probability of paper d being a member of D_{1t} and $P_{\theta_1}(c = 1|d, e, y = 1)$ is the probability of contribution of author e in paper d which has been written by at least one expert. In this learning model, we can use the members of set D_{1t} as well as D_{2t} to generate the training set. Each training instance can be described using the following set.

$$T_2 = \{(e, d, y, e_t)|d \in D_t, e \in A(d), e_t \in \{0, 1\}, y \in \{0, 1\}\}$$

Each member of T_2 (i.e. each training instance) has two labels. The y label indicates whether the paper d is a member of D_{1t}; if $y = 1$, then the label e_t indicates the relevance of author e in topic t. Given the relevance judgment e_t and the label y of each paper, the likelihood L' of the training data is as follows. We assume that the relevance judgments e_t are generated independently (we assume the same assumption for labels y).

$$L' = \prod_{m=1}^{|T_2|} \prod_{e \in A(d_m)} P_{\theta_1\theta_2}(c = 1, y = 1|d_m, e)^{y * e_t} P_{\theta_1\theta_2}(c = 0, y = 1|d_m, e)^{y(1-e_t)} P_{\theta_2}(y = 0|d_m)^{1-y} \quad (10)$$

Learning in this model is finding parameters θ_1 and θ_2 such that the above likelihood function is maximized. According to the definition of equation (9), we have the following

equations:

$$P_{\theta_1\theta_2}(c=1, y=1|d_m, e) = P_{\theta_2}(y=1|d_m)P_{\theta_1}(c=1|d_m, e, y=1)$$
$$P_{\theta_1\theta_2}(c=0, y=1|d_m, e) = P_{\theta_2}(y=1|d_m)(1 - P_{\theta_1}(c=1|d_m, e, y=1))$$
$$P_{\theta_2}(y=0|d_m) = 1 - P_{\theta_2}(y=1|d_m)$$

We model both $P_{\theta_2}(y = 1|d_m)$ and $P_{\theta_1}(c = 1|d_m, e, y = 1)$ by logistic functions on a linear combination of features. Formally, they are parameterized as follows:

$$P_{\theta_1}(c=1|d_m, e, y=1) = \sigma(\sum_{j=1}^{N_f} \beta_j f_j(d_m, e)) \quad (11)$$

$$P_{\theta_2}(y=1|d_m) = \sigma(\sum_{i=1}^{N_g} \alpha_i g_i(d_m)) \quad (12)$$

where N_f and N_g are the count of features for functions P_{θ_1} and P_{θ_2} respectively, and $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the standard logistic function. In above equations, α_i is the weight of the ith document feature $g_i(d)$ and β_j is the weight of the jth author-document feature $f_j(d, e)$. The weight parameters can be learned by maximizing the conditional log-likelihood of the data (i.e. equation (10)). Because there is no analytical solution, we use the BFGS Quasi-Newton for solving it. Table 1 indicates the features used for learning the contribution of authors in multi-author publications (i.e. $P_{\theta_1}(c = 1|d_m, e, y = 1)$). In this table, each feature is defined for a document d and one of its authors. As mentioned before, we categorize the features into three main groups. Structural features are defined based on the position of an author in the co-authorship network of the venue in which paper d is published in (i.e. $venue(d)$). For example, f_1 is the count of co-authors of an expert and f_2 is the PageRank value of an author in $venue(d)$. Moreover, f_3 is the AuthorRank [16] value, which is a modification of PageRank to measure the authority of authors in $venue(d)$. f_4 and f_5 are betweenness centrality measures (computed on weighted and un-weighted $venue(d)$) that represents the extent to which an author lies on the paths between other authors and can also be interpreted as measuring the influence an author has over the spread of information through the co-author network.
Temporal features are also defined for document and expert pairs, and represent the research longevity of an expert. Specifically, f_6 indicates the number of papers an expert e published before document d and f_7 is the overall research longevity of expert e. As the last feature group, activity-based features indicate the diversity and research quality of authors associated with a document. Specifically, f_8 is the number of different venues that an author is participated in. f_9 and f_{10} indicates the average and the sum of citation count of each author respectively. Finally, semantic relatedness feature (i.e. f_{11}) is the similarity score of the expert candidate's and the document's language models. The f_{11} is calculated based on the language models which mentioned in [4]. Features used for learning of $P_{\theta_2}(y = 1|d_m)$ should indicate the quality of paper d_m. We assume the quality of each paper is affected by the research quality of its authors. Therefore, we use the aggregate values of the features described in Table 1 to define document features. For example, features g_{i1} , g_{i2} and g_{i3} are the minimum, maximum and the average value of features f_i for all authors of the document d.

Table 1: Author-document association features organized in three groups

Structural	f_1	count of co-authors in venue of the associated document
	f_2	PageRank [7] in venue of the associated document
	f_3	AuthorRank [16] in venue of the associated document
	f_4	Un-weighted betweenness centrality [13] in venue of the associated document
	f_5	weighted betweenness centrality in venue of the associated document
Temporal	f_6	Count of papers previously published befor the associated document
	f_7	Research longevity (year)
Activity-based	f_8	count of attended venue
	f_9	average of citation count
	f_{10}	sum of citation count
	f_{11}	semantic relatedness [4]

The expert-document association probability $p(e|d)$ provides a ranking of candidates associated with a given document d, based on their contribution made to d. Intuitively, this probability is proportional to the prior probability of a candidate being expert(i.e. $p(e|d) \propto p(e)$). Therefore, in addition to the features that indicate the association strength of a document-expert pair, utilizing features that illustrate the expert authority is also reasonable.
Our proposed model is capable to assign different scores for different authors of a single documents. Therefore, considering each document as the relevant evidence, it is able to give more scores to more experienced candidates. After predicting the contribution probability for each author of the document d, we normalize the association probabilities such that $\sum_{e \in A(d)} p(e|d) = 1$.

5. EXPERIMENTS

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

- How **do** our learning models (i.e Model A and Model B) perform compared to each other?
- How do our expert finding model perform compared to the baseline models (i.e. the ULM and WLM approaches
- Whether our proposed model improves expert ranking independent of the baseline ranking models?
- What feature groups are likely more beneficial in terms of ranking expert candidates? Can integration of all
- How efficient is our proposed model in each topic?
- *How does our expert finding model rank top authors compared to state-of-the-art model*?

5.1 Experimental setup and metrics

In this section, we describe dataset and evaluation metrics.

5.1.1 Data Collection

We evaluate our models on the real-world DBLP bibliography data. Each DBLP record consists of several elements, such as "author", "title", "conference" and "date of publishing". Using these records, we could easily extract the co-author network and the mentioned features in Section 4.1. The citation and abstract information of papers were obtained from the Arnetminer [23] [3]. We get totally 1,632,442 papers and 1,033,321 authors in our data collection.

The evaluation of expert finding performance in such a large data collection is very challenging due to the scarcity of ground truth that can be examined publicly. Furthermore, it is impractical to obtain expert ratings for all authors. In order to measure the performance of the proposed methods, we use the benchmark dataset proposed by Deng et.al [10]. This dataset contains 17 queries [4] and a list of relevant experts for each topic that is manually created. This dataset cover both the broad and specific queries and has been widely used for evaluating expert finding task on bibliographic network [10, 11, 9]. In our experiments, we train our contribution learning model using all available topics rather than the given query.

5.1.2 Evaluation metrics

For the evaluation of task, several common IR metrics are employed in measuring the performance of our proposed models in different aspects. These metrics are precision at rank n ($p@n$), Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), $bpref$, and R-precision. $p@n$ measures the fraction of the top n retrieved results that are relevant experts for the given query which is defined as $p@n = \frac{\text{\# relevant experts in top n results}}{n}$

For a single query, AP is defined as the average of the $p@n$ values for all relevant documents as $AP = \frac{\sum_{n=1}^{N} p@n * rel(n)}{R}$ where n is the rank, N is the number of expert candidate retrieved, and $rel(n)$ is a binary function indicating the relevance of a given rank. AP emphasizes returning more relevant documents earlier, per query, and MAP is the mean value of the APs computed for all queries. The MRR is the average of the reciprocal ranks of results for a set of queries Q as $MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$. R-precision ($R - prec$) is defined as the precision at rank R where R is the number of relevant candidates for the given query. Besides the measurement of precisions, $bpref$ is a good score function that evaluates the performance from a different aspect, i.e., the number of non-relevant retrieved candidates. It is formulated as $bpref = \frac{1}{R}\sum_{r=1}^{N}(1 - \frac{\text{\# n ranked higher than r}}{R})$ where r is a relevant candidate and n is a member of the first R candidates judged non-relevant as retrieved by the system.

5.2 Experimental Results

In this subsection, we evaluate our proposed models for expertise ranking. First, we compare our two contribution learning models. Second, we compare our best proposed model with two state-of-the-art models (i.e. ULM and WLM). Then, we demonstrate which feature groups are likely more beneficial in terms of ranking expert candidates. Finally, we illustrate the efficiency of our proposed model for each topic, and top ranked authors compared to the best state-of-the-art model.

To demonstrate how the expertise ranking performance can be improved by our proposed approaches, we implemented ULM and WLM. As mentioned in Section 4, we proposed two contribution mining models (i.e. Model A and Model B). Table 2 indicates the performance measures introduced in Section 5.1.2 for *Model A* and *Model B* as well as the baseline methods. In this experiment, we follow the idea of WLM to define the prior relevance probability of a document proportional to the number of its citation. According to Table 2, *Model A(w)* improves all performance measures but slightly reduces the MRR measure. In contrast, *Model B(w)* improves all performance measures significantly. As mentioned before, *Model B* uses the training instances from both D_{1t} and D_{2t} sets, but *Model A* only uses the training instances from the D_{1t} set. Although, *Model A* can improve the expert ranking by estimating the contribution of authors, it is somehow biased for instances in set D_{1t} but the *Model B* utilizes all possible evidences to learn the contribution of authors. In the rest of the paper, we examine the properties of the *Model B* in comparison with the baseline models.

We compare the discriminative power of each feature group in recognizing the relevant expert for a given query. Table 3 indicates the performance measures of WLM and *Model B* which is trained on each feature group separately as well as the *Model B* which is trained by the all feature groups. According to this experiment, *structural*, *temporal* and *activity-based* feature groups can improve the WLM baseline ranking in terms of all evaluation measures. While *structural* and *activity-based* feature groups are more affective than the *temporal* feature group, they have almost the same discriminative power in recognizing relevant experts. Finally, learning the model by utilizing all of the feature groups can significantly improve expert ranking over state-of-the-art model in terms of all measures. Specifically, the proposed model can improve the $p@10$ up to 12.09%, the MAP up to 9.13% and the $bpref$ up to 14.00%.

Table 4 reports the evaluation results of our proposed models and two baseline models. In this experiment, we use two methods to determine the prior relevance probability of documents. In the first method, we use uniform prior probability for all documents (i.e. similar to the ULM approach) and in the second method, we use the weight proportional to the citation number of each document (i.e. similar to WLM approach). As it is clear in Table 4, the proposed model can improve the corresponding baseline models independent of the method used for estimating the document prior probabilities. This observation shows that, while the WLM approach focused on the document's quality for improving expert ranking, our method measures the contribution of each author of a document and gives more score to the leading expert candidates in a document.

Finally, we turn to a topic-level analysis of the comparisons illustrated in pervious experiments. We plot the differences in Average Precision between our proposed model and WLM (per query) in Figure 4. This figure shows that our proposed model is substantially preferable than the state-of-the-art model. According to Figure 4, it is obvious that our proposed model improves the average precision of WLM for

[3] http://arnetminer.org/lab-datasets/citation/

[4] this dataset can be accessed from http://www.cs.uiuc.edu/ hbdeng/data/queryset.txt

Table 2: Comparison of our two proposed models. In this experiment, w indicates that the prior relevance probability of a document is proportional to its number of citations and *all* indicates that all feature groups are utilized for learning proposed models. $*$ indicates the improvement is statistically significant ($\rho < 0.05$).

Method	p@10	p@20	p@30	MAP	MRR	bpref	R-prec
ULM	0.4647	0.3588	0.3118	0.3171	0.8188	0.3029	0.3576
WLM	0.5353	0.4265	0.3606	0.3865	0.9188	0.3571	0.4165
Model A(w-all)	0.5824	0.4441	0.3724	0.3971	0.8824	0.3800	0.4294
Improvement (%)	8.79%*	4.13%*	3.26%	2.73%*	-3.97%	6.41%*	3.10%
Model B(w-all)	**0.6000**	**0.4559**	**0.3982**	**0.4218**	**0.9412**	**0.4071**	**0.4506**
Improvement (%)	12.09%*	6.89%*	10.43%*	9.13%*	2.44%*	14.00%*	8.19%*

Table 3: Comparison of our proposed models with WLM for each feature group. $*$ indicates the improvement is statistically significant ($\rho < 0.05$).

Method	p@10	p@20	p@30	MAP	MRR	bpref	R-prec
WLM	0.5353	0.4265	0.3606	0.3865	0.9188	0.3571	0.4165
Model B(w-structural)	**0.6176**	**0.4588**	0.3959	0.4182	0.9312	0.4006	0.4447
Improvement (%)	15.37%*	7.57%*	9.79%*	8.20%*	1.35%	12.18%*	6.77%*
Model B(w-temporal)	0.5471	0.4294	0.3659	0.3924	0.9188	0.3624	0.4194
Improvement (%)	2.20%	0.68%	1.47%	1.53%*	0.00%	1.48%	0.70%
Model B(w-activity)	0.6059	0.4559	0.3965	0.4182	0.9312	0.4024	**0.4594**
Improvement (%)	13.19%*	6.89%*	9.96%*	8.20%*	1.35%	12.69%*	10.30%*
Model B(w-all)	0.6000	0.4559	**0.3982**	**0.4218**	**0.9412**	**0.4071**	0.4506
Improvement (%)	12.09%*	6.89%*	10.43%*	9.13%*	2.44%*	14.00%*	8.19%*

Table 4: comparison of model B with each baseline model. $*$ indicates the improvement is statistically significant ($\rho < 0.05$).

Method	p@10	MAP	MRR	bpref
ULM	0.4647	0.3171	0.8188	0.3029
Model B(u-all)	0.5353	0.3588	0.8941	0.3400
Improvement (%)	15.19%*	13.15%*	9.20%*	12.25%*
WLM	0.5353	0.3865	0.9188	0.3571
Model B(w-all)	**0.6000**	**0.4218**	**0.9412**	**0.4071**
Improvement (%)	12.09%*	9.13%*	2.44%*	14.00%*

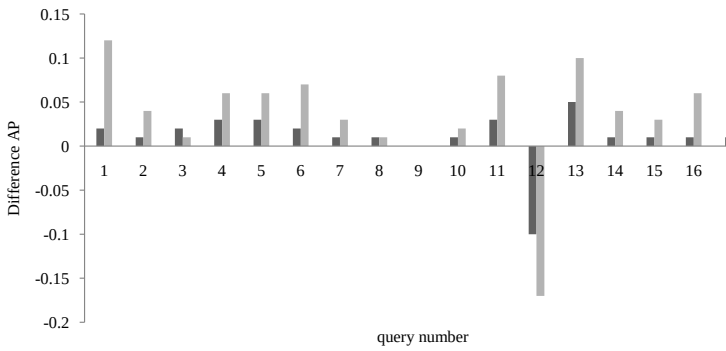

Figure 4: Topic-level differences in scores, WLM (baseline) vs. our proposed model.

all the queries except query 12. Interestingly, query 12 is "Language Model For Information Retrieval" topic, which is relatively a new field of research and therefore, most of its relevant experts are young researchers, which are less experienced than other authors. As mentioned before, our method tends to choose more experienced authors of a document as relevant experts. As a result, it is rational that our proposed model cannot improve average precision of query 12, but for all other queries, our model can improve the average precision of WLM.

To gain a better insight into our proposed model, we chose "Natural Language Processing" and "Information Retrieval" queries as the example cases to demonstrate the detailed results. The top-10 author lists ranked by two baseline models (i.e. ULM and WLM) and our proposed model are shown in Table 5, where the relevant experts are illustrated by underline. Undoubtedly, the results of our proposed method make more sense than baseline models. For example, considering the query "Natural Language Processing", we find that our proposed model can boost some relevant researchers such as "Sergei Nirenburg" and "Chris Mellish" into top 10. Similarly, results of "Information Retrieval" query show the same observation. This is because of distinguishing relevant expert author from other authors of a document. After looking into the details, we observe that our proposed model can successfully eliminate some irrelevant authors and return mostly relevant results.

6. CONCLUSION AND FUTURE WORK

In this paper, we studied the expertise ranking problem through learning contribution of each author in a scientific publication. We proposed two discriminative learning models to recognize the leading experts in a research group. We consider three effective feature groups to discriminate relevant experts from irrelevant ones. The experimental results show that our proposed approach can achieve significantly better results than the baseline and other state-of-the-art models. The proposed method in this paper can be used to

Table 5: The top-10 experts retrieved by two baseline models and our proposed model. Relevant experts are illustrated by underline.

ULM	WLM	*Model B*(w-all)
Natural Language Processing		
Barbara J. Grosz	Barbara J. Grosz	Barbara J. Grosz
Wolfgang Menzel	Wolfgang Menzel	Wolfgang Menzel
Alan F. Smeaton	Alan F. Smeaton	Yorick Wilks
Michael Hess	David D. Lewis	Alan F. Smeaton
Eva Hajičová	Ellen M. Voorhees	Sergei Nirenburg
Wlodek Zadrozny	Bonnie L. Webber	David D. Lewis
Bonnie L. Webber	Michael Hess	Bonnie L. Webber
Veronica Dahl	Eva Hajičová	Ellen M. Voorhees
Leonardo Lesmo	Wlodek Zadrozny	Raymond J. Mooney
Yorick Wilks	Raymond J. Mooney	Chris Mellish
Information Retrieval		
C. J. van Rijsbergen	C. J. van Rijsbergen	W. Bruce Croft
Gerard Salton	W. Bruce Croft	Gerard Salton
Norbert Fuhr	Gerard Salton	C. J. van Rijsbergen
W. Bruce Croft	Norbert Fuhr	Nicholas J. Belkin
Fabio Crestani	Nicholas J. Belkin	Norbert Fuhr
Nicholas J. Belkin	Fabio Crestani	Jian-Yun Nie
Jian-Yun Nie	Peter Ingwersen	Fabio Crestani
Mounia Lalmas	Mounia Lalmas	Ricardo A. Baeza-Yates
Thomas Mandl	James Allan	Mounia Lalmas
Giambattista Amati	Karen Sparck Jones	James Allan

recognize leading researchers in research communities, industrial groups, and also organizations. As a future work, we plan to implement our method in these environments.

7. REFERENCES

[1] A. M. Abbas. Weighted indices for evaluating the quality of research with multiple authorship. *Scientometrics*, 88(1):107–131, 2011.

[2] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Inf. Process. Manage.*, 45(1):1–19, 2009.

[3] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 551–558, 2007.

[4] K. Balog and M. De Rijke. Associating people and documents. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, pages 296–308, 2008.

[5] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Found. Trends Inf. Retr.*, 6:127–256, 2012.

[6] K. Balog, I. Soboroff, P. Thomas, N. Craswell, A. P. de Vries, and P. Bailey. Overview of the TREC 2008 enterprise track. In *The Seventeenth Text Retrieval Conference Proceedings (TREC 2008)*. NIST, 2009.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web*, pages 107–117, 1998.

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

[9] H. Deng, J. Han, M. R. Lyu, and I. King. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 71–80. ACM, 2012.

[10] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 163–172. IEEE Computer Society, 2008.

[11] H. Deng, I. King, and M. R. Lyu. Enhanced models for expertise retrieval using community-aware strategies. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(1):93–106, 2012.

[12] Y. Fang, L. Si, and A. P. Mathur. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 683–690. ACM, 2010.

[13] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, 1977.

[14] M. Karimzadehgan, R. W. White, and M. Richardson. Enhancing expert finding using organizational hierarchies. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 177–188. Springer-Verlag, 2009.

[15] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 467–476. ACM, 2009.

[16] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, 41(6):1462–1480, 2005.

[17] C. Macdonald and I. Ounis. Learning models for ranking aggregates. In *Proceedings of the 33rd European conference on Advances in information retrieval*, pages 517–529. Springer-Verlag, 2011.

[18] M. Neshati, H. Beigy, and D. Hiemistra. Multi aspect group formation using facility location analysis. In *Seventeenth Australasian Document Computing Symposium (ADCS 2012)*, pages 62–71, 2012.

[19] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 731–740. ACM, 2007.

[20] C. Sekercioglu. Quantifying co-author contributions. *Science*, 322:371, 2008.

[21] P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *Proceedings of 30th European conference on Advances in information retrieval*, pages 309–320. Springer-Verlag, 2008.

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

[23] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[24] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2), 2004.