Enhancing Expertise Retrieval Using Community-aware Strategies

Hongbo Deng
Dept. of CSE
The Chinese University of HK
Shatin, N.T., Hong Kong
hbdeng@cse.cuhk.edu.hk

Irwin King
Dept. of CSE,
The Chinese University of HK
Shatin, N.T., Hong Kong
king@cse.cuhk.edu.hk

Michael R. Lyu
Dept. of CSE
The Chinese University of HK
Shatin, N.T., Hong Kong
lyu@cse.cuhk.edu.hk

ABSTRACT

Expertise retrieval has received increased interests in recent years, whose task is to suggest people with relevant expertise. Motivated by the observation that communities could provide valuable insight and distinctive information, we investigate two community-aware strategies to enhance expertise retrieval. We first propose a new smoothing method using the community context instead of the whole collection for statistical language model in the document-based model. Furthermore, a query-sensitive AuthorRank is proposed to model the authors' authorities according to the community co-authorship networks, and then an adaptive ranking refinement method is developed to further enhance expertise retrieval. Experimental results demonstrate the effectiveness and robustness of both community-aware strategies.

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, language model, community, AuthorRank

1. INTRODUCTION

Expertise retrieval refers to the process of identifying a set of persons with relevant expertise for the given query. Traditionally, the expertise of a person is characterized based on the documents associated with the person. One of the state-of-the-art approaches [1, 3] is the document-based model using a statistical language model to rank experts. However, previous algorithms mainly consider the documents that are associated with the experts, while ignoring the community information that is affiliated with the documents and the experts. Therefore, it is essential to utilize the community-based information to enhance the expertise retrieval.

Given a set of documents and their authors, it is possible and often desirable to discover and infer the community information, in which contains a number of documents and authors for each community. Our approach is to deal with the expert-finding task in a real-world academic domain. Thus, it is reasonable to assume the academic communities have been formed automatically in the form of conferences and journals, where the researchers publish their papers, exchange their ideas, and co-author with each other.

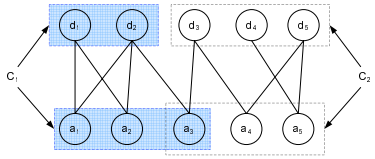


Figure 1: An example graph with two communities

We assume each document d_i can only belong to one community C_k, and each author a_j of the document is affiliated with the corresponding community C_k, so a single author may belong to multiple communities. An illustrated graph with two communities is sketched in Fig. 1. In this example, d_1 and d_2 as well as their associated authors form the community C_1, and meanwhile d_3, d_4 and d_5 as well as their authors form the community C_2. With such information, the community can be represented from two different perspectives, so as to obtain the community context (text information) based on the papers and the community co-authorship network based on the authors.

In this paper, we propose two community-aware strategies to enhance the expertise retrieval. The first one is the community-based smoothing method for statistical language model, which is employed to identify the most relevant documents so as to reflect the expertise retrieval in the document-based model. Moreover, the second strategy is developed to boost the document-based model using the community-sensitive authorities. More specifically, we propose a query-sensitive AuthorRank to model the authors' authorities based on the co-authorship networks, and develop an adaptive ranking refinement method to aggregate the ranking results. To illustrate our methodology, we apply the proposed methods to the expert finding task using the DBLP bibliography data. Experimental results demonstrate the effectiveness and robustness of the community-aware strategies.

2. RELATED WORK

Generally, there are two principal models for expertise retrieval: the candidate-based models [1, 2, 5] and the document-based models [1, 3, 5]. According to the comparison in [1], the document-based model could achieve better performance than the candidate-based model. Therefore, we choose the

CIKM'09, November 2–6, 2009, Hong Kong, China.

document-based model as the baseline, and propose several methods to further enhance this model.

This work is related to existing works in statistical language models [10, 11], which is employed to discover documents related to a query in the document-based model. Typically, a necessary step for the language model is to perform smoothing for the unseen query terms in the document, and several different smoothing methods have been proposed, such as Jelinek-Mercer smoothing and Bayesian smoothing using Dirichlet priors [11]. However, these smoothing methods only consider the collection as a whole, while our proposed smoothing method uses the community context information to smooth the language model.

Besides the categories described above, there are various methods proposed to extend or enhance the expertise retrieval. In [4], the authors propose a graph-based re-ranking model and apply it to expert finding for refining the ranking results. Furthermore, Karimzadehgan et al. [6] leverage the organizational hierarchy to enhance expert finding. Nevertheless, our proposed community-aware strategies are different from previous methods. In this work, We develop the query-sensitive AuthorRank as well as the adaptive ranking refinement strategy for the enhanced model.

3. MODELING EXPERTISE RETRIEVAL

3.1 Preliminaries

The problem of identifying experts is to estimate the probability of a candidate a_i being an expert given the query topic q. Using Bayes' theorem, the probability can be formulated as follows: $p(a_i|q) = \frac{p(a_i,q)}{p(q)} \propto p(a_i, q)$. The probability $p(q)$ is a constant, so it can be ignored for ranking purposes. To derive $p(a_i|q)$, it is equivalent to estimate the joint probability $p(a_i, q)$. In [3], Deng et al. propose a document-based model to aggregate the expertise of an expert according to the relevance and importance of the associated documents. The joint probability can be decomposed as

$$p_d(a_i, q) = \sum_{d_j \in D} p(d_j)p(q|d_j)p(a_i|d_j), \tag{1}$$

where $p(d_j)$ is the prior probability of a document, $p(q|d_j)$ means the relevance between q and d_j, and $p(a_i|d_j)$ represents the association between the candidates and the documents. Under this model, $p(d_j)$ is estimated based on the citation of the document $N_c(d)$: $p(d) \propto \log(10 + N_c(d))$. Suppose a document has multiple authors in total, each author is assumed to have the same knowledge about the topics described in the document,

$$p(a|d) = \begin{cases} \frac{1}{N_a(d)}, & (a \text{ is the author of } d) \\ 0, & (\text{otherwise}) \end{cases} \tag{2}$$

where $N_a(d)$ is the number of authors for the document.

3.2 Smoothing Using Community Context

In the document-based model, one of the key challenges is to determine the probability of a query given a document $p(q|d)$. According to the statistical language model, we infer a document language model θ_d for each document. The relevance score of document d with respect to query q is then defined as the conditional probability $p(q|\theta_d)$,

$$p(q|\theta_d) = \prod_{t_i \in q} p(t_i|\theta_d), \tag{3}$$

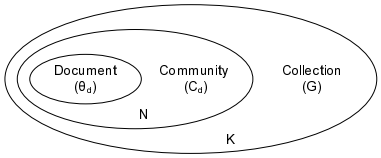


Figure 2: A graph representation of the relationships between documents, communities and the entire collection.

where $p(t|\theta_d)$ represents the maximum likelihood estimator of the word in a document d. In order to assign nonzero probabilities to unseen words, it is important to incorporate the smoothing methods in estimating the document language model.

In general, each word is smoothed by the same collection language model. However, the community provides distinctive information for its documents. Figure 2 illustrates the relationships between the documents, the communities and the whole collection. Basically, a document will somewhat share much more common information with its community C_d rather than the whole collection G. Therefore, it would be more reasonable to employ the distinctive community language model, instead of the whole collection language model, to smooth different document models. The community language model is defined as

$$p(t|C_d) = \frac{\sum_{d_j \in C_d} n(t, d_j)}{\sum_{d_j \in C_d} |d_j|}. \tag{4}$$

One popular way to smooth the maximum likelihood estimator is the Jelinek-Mercer smoothing method:

$$p(t_i|\theta_d) = (1 - \lambda)\frac{n(t_i, d)}{|d|} + \lambda p(t_i|C_d), \tag{5}$$

where λ is the parameter to control the amount of smoothing, $n(t_i, d)$ is the count of word t_i in the document d, and $|d|$ is the number of the words in d. Accordingly, the community-smoothed language model is obtained

$$p(q|\theta_d) = \prod_{t_i \in q} \left((1 - \lambda)\frac{n(t_i, d)}{|d|} + \lambda p(t_i|C_d) \right). \tag{6}$$

Note here the document d belongs to the community C_d. So far, there are two different language models, i.e., the collection-smoothed language model (baseline model) and the community-smoothed language model, for calculating $p(q|\theta_{d_j})$. Therefore, two different models can be combined as shown in the upper part of Table 1.

4. ENHANCED MODELS WITH COMMUNITY-AWARE AUTHORITIES

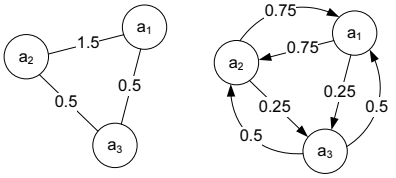
4.1 Discovering Authorities in a Community

In a community, the authors' relationships can be described using a co-authorship network, which is an important category of social networks [8].To quantify the edge weight, the co-authorship frequency [7] is proposed as the sum of values for all papers co-authored by a_i and a_j, $f_{ij} = \sum_{k=1}^{N} \frac{\delta_i^k \delta_j^k}{n_k - 1}$, where $\delta_i^k = 1$ if a_i is one of the authors of the paper d_k, otherwise $\delta_i^k = 0$, and n_k is the number of authors in paper d_k. This gives more weight to authors who

Table 1: Combination of different methods.

Model	c[a]	E[b]	Remarks
$DM(w)$	0	0	baseline model
$DM(wc)$	1	0	community-based smoothing
$EDM(w)$	0	1	enhanced $DM(w)$
$EDM(wc)$	1	1	enhanced $DM(wc)$

[a] smoothing using community (1) or collection (0)
[b] enhancing with community-aware authorities (1) or no enhancement (0)



(a) (b)

Figure 3: Co-authorship graph with: (a) co-authorship frequency, and (b) normalized weight.

co-publish more papers together. Let us take the community C_1 in the Fig. 1 as an example, the graph with the co-authorship frequency is illustrated in Fig. 3(a). In general, the link weight w_{ij} from a_i to a_j is defined by normalizing the co-authorship frequency from a_i as $w_{ij} = \frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}$. This normalization ensures that the weights of an author's relationships sum to one, as shown in Fig. 3(b) for C_1.

For each community, a weighted co-authorship graph can be easily built. We therefore utilize AuthorRank [7], a modification of PageRank [9], to measure the authority for the authors within this community as

$$p(a_i|C_k) = (1-\alpha)\frac{1}{N_a(C_k)} + \alpha \sum_{j=1}^{N_a(C_k)} w_{ij} \cdot p(a_j|C_k), \quad (7)$$

where $N_a(C_k)$ is the number of authors in the community C_k, and $p(a_i|C_k)$ is the authority (i.e., AuthorRank) of the author a_i satisfying $\sum_i p(a_i|C_k) = 1$.

4.2 Community-Sensitive AuthorRank

The AuthorRank described above calculates the authorities for the authors within a community, but it is independent of any particular query topic. To identify a set of experts for a given query, we propose a community-sensitive AuthorRank to generate query-specific authority scores for authors at query time.

Given a query q, we compute the probability for each community C_k the following:

$$p(C_k|q) = \frac{p(C_k) \cdot p(q|C_k)}{p(q)} \propto p(C_k) \prod_{t_i \in q} p(t_i|C_k), \quad (8)$$

where $p(t_i|C_k)$ is easily computed from the community language model as Eq. (4). The quantity $p(C_k)$ is not as straightforward. We model it as related to the number of authors $N_a(C_k)$ and the average citation per paper $N_c(C_k)$ in the community C_k; that is

$$p(C_k) \propto N_a(C_k) \cdot \log(10 + N_c(C_k)). \quad (9)$$

The underlying idea is that the community prior is proportional to the size and quality of the community.

According to Eq. (8), we retrieve top-k communities that are highly related to the query. Finally, we compute the query-sensitive authority score for each author as follows,

$$p_c(a_i|q) = \sum_k p(C_k|q)p(a_i|C_k), \quad (10)$$
$$\propto \sum_k p(C_k)p(q|C_k)p(a_i|C_k).$$

The authors are ranked according to the composite score $p_c(a_i|q)$. The above community-sensitive AuthorRank has the following probabilistic interpretation. Suppose C_k be a "virtual" document, it becomes the document-based model as Eq. (1). Thus the community-sensitive AuthorRank can be regarded as a high-level document-based model that captures the high-level and general aspects for a given query.

4.3 Ranking Refinement Strategy

Based on the document-based model and the community-sensitive AuthorRank (i.e., community-based model), we obtain two kinds of ranking results $\vec{Rd}$ and $\vec{Rc}$, which reflect the authors' expertise from different perspectives. The ranking list $\vec{Rd}$ captures more specific and detailed aspects matching with the given query, as it measures the contribution of each document individually. In contrast, the ranking list $\vec{Rc}$ reflects more general and abstract aspects matching with the given query. Therefore, we consider the ranking refinement strategy by leveraging the community-sensitive AuthorRank to boost the document-based model.

If the community-sensitive AuthorRank could retrieve many common authors within the top-k results as identified by the document-based model, the community-sensitive AuthorRank may contribute a lot to refine the document-based model; otherwise vice versa. Based on this scheme, we utilize the Jaccard coefficient to measure the similarity between two top-k ranking results, which is defined as $J = \frac{|\vec{Rd} \bigcap \vec{Rc}|}{|\vec{Rd} \bigcup \vec{Rc}|}$. Then we adopt this measurement for an adaptive ranking refinement as follows,

$$S(a_i) = \frac{1}{Rd(a_i)} + \delta(a_i) \cdot J \cdot \frac{1}{\hat{Rc}(a_i)}, \quad (11)$$

where $Rd(a_i)$ be the rank of author a_i in $\vec{Rd}$, $\hat{Rc}(a_i)$ be the rank of author a_i in $\hat{Rc}$ (i.e., $\vec{Rd} \bigcap \vec{Rc}$), and $\delta(a_i) = 1$ if a_i is one of the intersected authors, otherwise $\delta(a_i) = 0$. The intuition behind this method is that the authors, which are identified in both $\vec{Rd}$ and $\vec{Rc}$, should be boosted ahead in the ranking results. The final results are ranked according to the refined score $S(a_i)$. By applying the ranking refinement strategy to the previous two different document-based models, we obtain two enhanced models as shown in Table 1.

5. EXPERIMENTAL EVALUATION

We evaluate the performance of our proposed models with different settings. In this section, we first introduce the experimental setup, and then present the experimental results.

5.1 Experimental Setup

The dataset that we study is the DBLP bibliography data, which contains over 1,100,000 XML records as of March 2009. In summary, the data collection for experiments include 1,184,678 papers, 696,739 authors, and 3,143 communities. In order to measure the performance of our proposed

Table 2: Comparison of different methods.

Method	P@10	P@20	R-prec	MAP
DM(w)	0.688	0.503	0.485	0.363
DM(wc)	0.688	0.527	0.494	0.377
EDM(w)	0.706	0.55	0.532	0.403
EDM(wc)	**0.712**	**0.568**	**0.533**	**0.409**
DM(wc)/DM(w)	0%	+4.7%	+2.0%	+3.8%
EDM(w)/DM(w)	+2.6%	+9.4%	+9.8%	+10.9%
EDM(wc)/DM(wc)	+3.4%	+7.8%	+7.9%	+8.4%

methods, we manually created the ground truth because of the scarcity of such data that can be examined publicly. For each query, a list of experts is collected through the method of pooled relevance judgments with human assessment efforts. The benchmark dataset used for the evaluation contains 17 query topics and 17 expert lists. In our experiments, we report the results of P@10, P@20, R-prec, and MAP.

5.2 Experimental Results

5.2.1 Comparison of Different Models

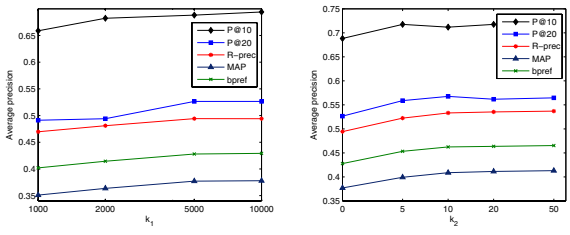
To validate the effect of the community-based smoothing method, we evaluate and compare the performance of four different methods. In Table 2, the first part shows the absolute precisions of these methods, and the second part illustrates the percentages of relevant improvements.

According to the first part, it is obvious that $EDM(wc)$ achieves the best performance in all the metrics, such as 0.568 for P@20 and 0.409 for MAP. When looking at the relative improvements, we can see that $DM(wc)$ improves over $DM(w)$ from 2.0% to 4.7% in most metrics besides P@10. This is because the smoothing method using community context can boost the performance of the language model. As expected, the enhanced models $EDM(w)$ and $EDM(wc)$ perform better than their corresponding document-based models $DM(w)$ and $DM(wc)$, respectively. For MAP metric, we can see that $EDM(wc)$ improves over $DM(wc)$ by 8.4%, and $EDM(w)$ over $DM(w)$ by 10.9%. In terms of the comparisons using other metrics, we observe similar substantial improvements. All the experimental results demonstrate the effectiveness of the enhanced model, which could further boost the performance of document-based models.

5.2.2 Discussion and Detailed Results

As mentioned before, we only retrieve the top-k_1 relevant documents for the document-based model, and identify top-k_2 relevant communities for the community-sensitive AuthorRank as well. The parameters k_1 and k_2 used in previous subsections are set to 5,000 and 10, individually. To investigate the effect of these two parameters, we designed the following experiments.

To examine the effect of k_1, we choose the document-based model $DM(wc)$, and evaluate it with 4 different values (from 1,000 to 10,000). The experimental results for different k_1 are illustrated in Fig. 4(a). In this figure, we can see the performance becomes better for greater k_1 used in the document-based model. We believe the reason is that more documents can better capture the complete expertise. However, larger k_1 may result in longer processing time. Therefore, a good tradeoff is to set $k_1 = 5000$. To investigate the effect of k_2, we fix $k_1 = 5000$, and choose to compare the model $EDM(wc)$ with several different values from 0 to 50.



(a) $DM(wc)$ (b) $EDM(wc)$

Figure 4: The effect of varying the parameters (k_1 and k_2) in (a) the document-based model $DM(wc)$ and (b) the enhanced model $EDM(wc)$.

Here, $k_2 = 0$ in $EDM(wc)$ represents its document-based model $DM(wc)$. As shown in Fig. 4(b), when incorporating the community-sensitive AuthorRank in the enhanced model ($k_2 > 0$), the performance is improved compared to the document-based model ($k_2 = 0$). The precisions first increase then level off as k_2 grows. In general, the enhanced model $EDM(wc)$ is relatively robust for different k_2, and achieves good results when $k_2 = 10$.

6. CONCLUSIONS

In this paper we present the community-aware strategies for enhancing expertise retrieval, including the new smoothing method with the community context and the community-sensitive AuthorRank based on the co-authorship networks. Extensive experiments show that the improvements of our enhanced models are significant and consistent.

7. ACKNOWLEDGMENTS

This work is supported by two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK4128/08E and Project No. CUHK4158/08E).

8. REFERENCES

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

[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] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *ICDM*, pages 163–172, 2008.

[4] H. Deng, M. R. Lyu, and I. King. Effective latent space graph-based re-ranking model with global consistency. In *WSDM*, pages 212–221, 2009.

[5] H. Fang and C. Zhai. Probabilistic models for expert finding. In *ECIR*, pages 418–430, 2007.

[6] M. Karimzadehgan, R. W. White, and M. Richardson. Enhancing expert finding using organizational hierarchies. In *ECIR*, pages 177–188, 2009.

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

[8] M. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101(suppl 1), 2004.

[9] L. Page and S. Brin. The anatomy of a large-scale hypertextual web search engine. In *WWW*, 1998.

[10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[11] C. Zhai. Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–215, 2008.