BMExpert: Mining MEDLINE for Finding Experts in Biomedical Domains Based on Language Model

Beichen Wang, Xiaodong Chen, Hiroshi Mamitsuka, and Shanfeng Zhu

Abstract—With the rapid development of biomedical sciences, a great number of documents have been published to report new scientific findings and advance the process of knowledge discovery. By the end of 2013, the largest biomedical literature database, MEDLINE, has indexed over 23 million abstracts. It is thus not easy for scientific professionals to find experts on a certain topic in the biomedical domain. In contrast to the existing services that use some *ad hoc* approaches, we developed a novel solution to biomedical expert finding, BMExpert, based on the language model. For finding biomedical experts, who are the most relevant to a specific topic query, BMExpert mines MEDLINE documents by considering three important factors: relevance of documents to the query topic, importance of documents, and associations between documents and experts. The performance of BMExpert was evaluated on a benchmark dataset, which was built by collecting the program committee members of ISMB in the past three years (2012-2014) on 14 different topics. Experimental results show that BMExpert outperformed three existing biomedical expert finding services: JANE, GoPubMed, and eTBLAST, with respect to both MAP (mean average precision) and P@50 (Precision). BMExpert is freely accessed at http://datamining-iip.fudan.edu.cn/service/BMExpert/.

Index Terms—Biomedical text mining, expert finding, language model, information retrieval.

Introduction

As a fast growing discipline, biomedical science generates and validates many new scientific hypotheses. New knowledge and findings are constantly reported. By the end of 2013, the largest biomedical literature database, MEDLINE[1], has indexed over 23 million citations (abstracts), most of which were published after 1946. In the last few years, the number of indexed citations in MEDLINE has increased exponentially. In 2011, 2012, and 2013, for example, the number of citations added in MEDLINE was 758,918, 777,559, and 868,372, respectively. Obviously, it is not an easy task for scientific professionals to find relevant information from MEDLINE. To alleviate this problem, a web-based tool, PubMed [1], is developed by the NLM National Center for Biotechnology Information (NCBI) for providing a searching service on finding relevant citations for a given query, i.e. keywords. PubMed uses a boolean model to retrieve relevant citations, which are then sorted by publication date, title, the position in the author list (first or last author) or relevance and finally displayed for users. To enhance the quality and visualization of literature searching results by PubMed, several web servers, including HubMed [2], GoPubMed [3] and XploreMed [4] have been developed. Moreover, a number of clustering algorithms have been proposed to help users to navigate and locate some interesting documents [5]–[10].

In addition to retrieving and clustering biomedical documents, finding experts on a specific topic in biomedical domain is very important for biomedical researchers. Many techniques and special terms have been emerging in biomedical science. Typical examples are Induced Pluripotent Stem (iPS) cells, personalized medicine and translational research, etc. To grasp such new technology accurately and quickly, a key step is to find experts of the focused field and read papers by experts [11]. Moreover, finding experts would be useful in many other applications, such as finding appropriate paper reviewers or committee members for funding evaluations or organizing conferences.

Although there are several web servers that can be used to find biomedical experts, such as GoPubMed [3], eTBLAST [12] and Jane [13], the underlying principles behind these services have three main weaknesses. Firstly, their main purpose is not about finding experts with respect to a specific topic query. A typical input to Jane and eTBLAST is the title or abstract of a query paper (though keyword based queried are possible). Jane and eTBLAST are then to suggest similar publications of the input and relevant journals as well as experts. As an alternative of PubMed, GoPubMed can find relevant search results faster than PubMed by using knowledge-based search with Gene Ontology (GO) and Medical Subject Heading (MeSH). Secondly, only partial information is considered in these services for finding experts. To score an expert with respect to a topic query, we usually first retrieve all documents

- *Beichen Wang, Xiaodong Chen and Shanfeng Zhu are with the School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China. E-mail: zhusf@fudan.edu.cn*
- *Hiroshi Mamitsuka is with Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan. E-mail: mami@kuicr.kyoto-u.ac.jp*

1. http://www.nlm.nih.gov/pubs/factsheets/medline.html

written by the expert. We then score every document by considering three important factors: the relevance of the document to the query, the importance of the document and the association between the document and the expert. However, none of GoPubMed, eTBLAST and Jane has considered the importance of documents. In addition, for the paper with multiple authors, both GoPubMed and Jane treat all authors equally, which is not necessarily reasonable for biomedical document with many authors. Thirdly, the scoring methods of these services for finding experts are pretty much *ad hoc*, which lacks theoretical justification. Given a specific query, GoPubMed extracts authors of the papers relevant to the query and sorts the authors according to the number of the related papers. In this case, all retrieved documents have the same relevance score of 1. In contrast, the document-query relevance scores in eTBLAST are derived from sentence alignment, and those in Jane directly from Lucene [14]. With these weaknesses, it is not surprising that these services can hardly achieve satisfied performance for finding experts in biomedical domain. In addition, PubFocus, a citation analysis software, had a function of ranking biomedical experts, but currently cannot provide any on-line service for finding experts [15].

To overcome the shortcomings of existing services, we developed a novel biomedical expert finding service, BMExpert, based on the language model, simultaneously considering three important factors in a uniform framework, the relevance of a document to a topic query, the importance of a document and the association between a document and an expert. The performance of BMExpert was evaluated on a benchmark dataset, which was built by using the program committee members of ISMB (Annual International Conference on Intelligent Systems for Molecular Biology: 2012-2014) with respect to 14 different topics. Experimental results show that BMExpert outperforms all existing biomedical expert finding services, such as Jane, GoPubMed, and eTBLAST, in both MAP (mean average precision) and P@50. We further examined the effect of different configurations of three important factors in BMExpert.

The rest of this paper is organized as follows: Section 2 reports the underlying algorithms of BMExpert. Section 3 shows the experimental results and the performance comparisons of BMExpert with three existing biomedical expert finding services, Jane, GoPubMed and eTBLAST. Finally we discuss the results and conclude in Section 4.

Methods

Preliminary

Given a topic query q, an expert candidate e and a document set D_e, which consists of all documents written by candidate e, $S(e,q)$ is the score of candidate e with respect to topic q. Expert finding can be defined as a problem of generating a ranked list of experts on a given topic.

BMExpert

In contrast to existing solutions of biomedical expert finding using some *ad hoc* approaches, BMExpert is developed according to a weighted language model, which has been used for expert finding in computer science [16], [17]. Given a topic, there must be many relevant papers, where the paper authors are all possible candidates of experts. The main idea of a language model for expert finding is to estimate the probability of a candidate e being expert on topic q, denoted by $p(e|q)$. That is, we use $p(e|q)$ to represent $S(e,q)$, the score of e with respect to q. $p(e|q)$ is given as follows:

$$p(e|q) = p(e,q)/p(q)$$

Here $p(e,q)$ is the joint probability of query topic q and candidate e, and $p(q)$ is the probability of query topic q. Since $p(q)$ is a constant for a specific query topic, we can ignore this constant. Thus the probability of candidate e being an expert on given query topic q is proportional to the joint probability of query topic q and candidate e:

$$p(e|q) \propto p(e,q)$$

In order to estimate $p(e,q)$, we introduce a variable d to represent document (or paper).

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

Given d, we assume that the probability of observing q is independent from observing e. Then we have

$$\begin{aligned} p(e,q) &= \sum_{d \in D_e} p(e,q,d) = \sum_{d \in D_e} p(d) \times p(e,q|d) \\ &= \sum_{d \in D_e} p(d) \times p(q|d) \times p(e|d) \end{aligned} \tag{1}$$

Here we can see that $p(e,q,d)$ is determined by three factors, $p(d)$, $p(q|d)$ and $p(e|d)$.

1) $p(d)$
The $p(d)$ in Eq.(1) is the prior probability of document d, corresponding to the importance of paper d that can be rewritten as a weight factor w_d.

2) $p(e|d)$
The $p(e|d)$ in Eq.(1) is a candidate weight, namely, the probability of candidate e given document d.

3) $p(q|d)$
We compute $p(q|d)$ according to the language model. The $p(q|d)$ can be written as $p(q|\theta_d)$, since this probability has a certain structure further as follows:

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)$$

where term t occurs in query q. To estimate $p(t|\theta_d)$, we can use a smoothing method between two probabilities, $p(t|d)$ and $p(t)$ as follows:

$$p(t|\theta_d) = (1-\lambda)p(t|d) + \lambda p(t)$$

where $p(t)$ is the probability of term t occurring in all documents relevant to query q, and $p(t|d)$ is the probability of term t occurring only in document

d. We note that the smoothing factor λ takes a value between zero and one, meaning that if λ is closer to one, $p(t|\theta_d)$ can be more uniform.

The final probabilistic structure of the weighted language model for BMExpert is given as follows:

$$\begin{aligned} p(e,q) &= \sum_{d\in D_e} p(d)\times p(q|d)\times p(e|d) \\ &= \sum_{d\in D_e} p(d)\times \prod_{t\in q}((1-\lambda)p(t|d)+\lambda p(t))\times p(e|d) \end{aligned} \quad (2)$$

Generalization

In the weighted language model for BMExpert, we score a candidate e with respect to topic q in a probabilistic way, and $S(e,q)$ is computed by the join probability of e and q, $p(e,q)$. By introducing a variable d to represent a paper written by e, the score of a candidate e on topic q is determined by three factors, $p(d)$, $p(q|d)$ and $p(e|d)$. In fact, the idea is intuitive. For scoring expert e with respect to q, $S(e,q)$, it is natural to add up the score of each document d written by expert e with respect to topic q. That is,

$$S(e,q)=\sum_{d\in D_e} s(d,q)$$

To compute $S(d,q)$, we need to consider three factors: $R(d,q)$, the relevance of d to q, $I(d)$, the importance of d and $A(d,e)$, the association of e with d. $R(d,q)$ is the degree of relevance between document d and topic q. $I(d)$ measures the importance of document d, which weighs more on those in the high profile journal or have received many citations. $A(d,e)$ reflects the associations between expert e and document d, which is especially useful for documents with multiple authors. An intuitive approach to computing $S(d,q)$ is to integrate all three factors together. That is,

$$s(d,q)=R(d,q)\times I(d)\times A(d,e)$$

Overall, we have

$$S(e,q)=\sum_{d\in D_e} s(d,q)=\sum_{d\in D_e} R(d,q)\times I(d)\times A(d,e) \quad (3)$$

This is a generalized document based framework for expert finding (GDFEF). In BMExpert, $I(d)$ is computed by $p(d)$, $A(d,e)$ by $p(e|d)$, and $R(d,q)$ by $p(q|d)$ according to the language model. In this sense, the weighted language model for expert finding implements GDFEF in a probabilistic way.

GoPubMed, Jane and eTBLAST

Here we analyze the scoring approaches used in GoPubMed, Jane and eTBLAST by using GDFEF.

1) GoPubMed

GoPubMed does not consider the importance of the document, and treats all authors of each document equally. That is, $I(d)=A(d,e)=1$. Given a query, GoPubMed relies on PubMed to retrieve relevant documents. It treats all retrieved documents equally, and sets the same relevance scores 1 to all of them. That is,

$$S_{GoPubMed}(e,q)=|\{d|d\in D_{e,q}\}| \quad (4)$$

where $D_{e,q}$ is the set of retrieved documents written by e with respect to q.

2) Jane

Jane is the same as GoPubMed. That is, $I(d)=A(d,e)=1$. It uses Lucene to compute the relevance score between q and d. That is,

$$S_{Jane}(e,q)=\sum_{d\in D_e} R_{Lucene}(d,q) \quad (5)$$

where $R_{Lucene}(d,q)$ is the relevance score between d and q by Lucene.

3) eTBLAST

eTBLAST does not consider the importance of document, either. It uses sentence alignment to score the document with respect to a topic query. For measuring the associations between expert and documents, it uses a simple weighting scheme: +3 for last authors, +2 for first authors and +1 for all other authors. That is,

$$S_{eTBLAST}(e,q)=\sum_{d\in D_e} R_{align}(d,q)\times A_w(d,e) \quad (6)$$

where $R_{align}(d,q)$ is the sentence alignment score between d and q, and $A_w(d,e)$ is the weighting scheme described above.

Overall, we can see that existing biomedical expert services only consider a subset of three important factors in some *ad hoc* manners, while BMExpert considers all of them probabilistically under a formal framework of the language model.

Experimental Results

Benchmark Dataset

It is a very hard question to find an expert on a given research topic, because researchers may have their own opinions even if they are asked to select top experts in their field. This point makes it difficult to generate a benchmark dataset for evaluating BMExpert. To reduce possible bias and make the comparison fair, we used all ISMB program committee members (including area chairs) from 2012 to 2014 as a gold standard dataset. This dataset contains 14 topics and the number of experts on each topic varies from 16 to 163. These 14 topics are "Applied Bioinformatics","Bioimaging and Data Visualization", "Databases and Ontologies", "Disease Models and Epidemiology", "Evolution and Comparative Genomics", "Gene Regulation and Transcriptomics", "Mass Spectrometry and Proteomics", "Metabolic Networks", "Population Genomics", "Protein Interactions and Molecular Networks", "Protein Structure

TABLE 1
Differences between BMExpert and other three biomedical expert finding services: GoPubMed, eTBLAST and Jane

Server	Purpose	Considering factors		
		Importance of paper	Relevance between query and paper	Association between expert and paper
BMExpert	Find experts for a specific topic query	Recentness, SCI impact factors, Domain	Score of top 2000 papers using language model	Options for considering first author, last author or all authors
GoPubMed	Explore PubMed with Gene Ontology	None	None	Equal for each author
eTBLAST	Identify expert reviewers, appropriate journals and similar publications.	None	Scores of top 400 papers using sentence-alignment	+3 for last authors, +2 for first authors and +1 for all other authors
Jane	Suggest journals and find reviewers	None	Scores of top 50 papers using Lucene	Equal for each author

and Function", "RNA Bioinformatics", "Sequence Analysis", and "Text Mining". (Please see Supplementary materials for the list of experts on each topic). It should be noted that there may have many other experts of these fourteen topics. This means that the performance of BMExpert and other compared methods might be underestimated. Nevertheless, these researchers must have expertise in the corresponding areas. Thus we emphasize that the benchmark dataset we used provides a reasonable platform to compare the performance of different methods.

Evaluation Metrics

We used two criteria for evaluation after obtaining top n experts from each of the competing methods: $P@n$ and AP (average precision) [18].

$P@n$ is obtained by dividing the number of correct outputs (i.e. true experts) in the top n outputs by n. $P@n$ is also called "precision". In our experiments, we use $n = 50$, denoted as $P@50$.

For a query, we can compute average precision (AP) as follows:

$$AP = \frac{\sum_{i=1}^{n} p@i \times rel_i}{N_r}$$

where N_r is the total number of true experts, $P@i$ is the precision at the top i ($< n$) outputs, and rel_i is 1 if the result at position i is correct (i.e. an expert), otherwise zero. Note that for the experts who do not appear in the top n results (here we use $n = 100$), the precision will be set to 0. We can see that AP favors the method that returns experts in higher (upper or closer to top) positions.

Experimental settings

We compared the performance of BMExpert with three existing biomedical expert services: GoPubMed, eTBLAST and Jane. Considering the characteristics of biomedical domain, we examined various settings of computing document importance ($p(d)$ or w_d) and document-expert associations ($p(e|d)$) in BMExpert. We examined the effect of using the impact factor to estimate the importance of documents. Impact factor is computed from the number of citations, which is an important criterion in biomedical domain for paper evaluation. If a paper was published in 2014, we used the latest five-year impact factor as an approximation until the new impact factor is available. Specifically, w_d is computed as $w_d = \ln(e + IF)$, where IF is the SCI (Science Citation Index) impact factor of the journal, in which paper d appears. Since the benchmark dataset was built from bioinformatics domain, we selected a small set of journals that are closely related to bioinformatics (See the supplemental material for the journal list). We then examined the effect of BMExpert by restricting the paper published in these journals. In addition, we studied the effect of restricting the papers published in the last few years. This is reasonable because the users may want to find experts who have published relevant papers "recently". For the effect of setting document-expert associations, since many biomedical papers have a large number of authors, we considered five options, 1) all authors equally, 2) only the first author, 3) only the last author, 4) the first and last authors and 5) all authors with eTBLAST weighting scheme. For computing the relevance between document and query $p(d|q)$, according to [18], we set the smoothing parameter $\lambda = 0.6$. Table 1 summarizes the configuration of BMExpert, Jane, GoPubMed and eTBLAST, respectively, and the settings of Jane, GoPubMed and eTBLAST are from their published papers. For the performance comparison of different methods over benchmark dataset, all the servers were accessed on the same date: July 15, 2014.

Results

We first examined the performance of BMExpert with a general setting. In this case, we did not use the publication date and journal name to restrict the retrieved papers. We treated all papers equally without considering the impact factor. For the document-expert associations, we considered only the first and last authors which would be a reasonable manner for handling biomedical publications with many authors. Table 2 illustrates the performance comparisons of BMExpert with Jane, GoPubMed and eTBLAST. BMExpert achieved the highest average P@50 of 6.71%, followed by GoPubMed (3.86%), Jane (3.14%) and eTBLAST (2.14%), respectively. Interestingly, for two "difficult" topics, "Disease Models and Epidemiology" and

TABLE 2
The performance comparison of BMExpert with default settings against Jane, GoPubMed and eTBLAST (%)

Topics	P@50				AP			
	BMExpert	Jane	GoPubMed	eTBLAST	BMExpert	Jane	GoPubMed	eTBLAST
Applied Bioinformatics	2	2	**6**	2	0.06	0.06	**0.77**	0.01
Bioimaging and Data Visualization	**14**	0	2	4	**4.11**	0.05	0.12	0.5
Databases and Ontologies	2	4	**8**	0	0.21	1.02	**1.39**	0.04
Disease Models and Epidemiology	0	0	0	0	0	0	0	0
Evolution and Comparative Genomics	**2**	0	0	0	**0.18**	0	0	0
Gene Regulation and Transcriptomics	0	**2**	0	0	0	**0.04**	0	0
Mass Spectrometry and Proteomics	0	**2**	**2**	**2**	0	**1.33**	0.11	0.38
Metabolic Networks	**16**	6	8	2	**14.27**	2.52	3.11	0.65
Population Genomics	0	0	0	0	0	0	0	0
Protein Interactions and Molecular Networks	**12**	4	10	8	2.29	0.41	1.22	**4.77**
Protein Structure and Function	**16**	6	2	0	**3.46**	0.49	0.32	0
RNA Bioinformatics	**14**	6	12	0	**16.59**	8.03	13.53	0
Sequence Analysis	0	**2**	0	0	0	**0.08**	0	0
Text Mining	**16**	10	4	12	16.77	13.71	2.77	**17.85**
Average	**6.71**	3.14	3.86	2.14	**4.14**	1.96	1.68	1.73

TABLE 3
The performances of BMExpert with incorporating paper recentness (paper published in the last x years: one, three, five, ten and all)(%).

Topics	P@50					AP				
	one	three	five	ten	all	one	three	five	ten	all
Applied Bioinformatics	**4**	**4**	**4**	2	2	0.09	0.09	0.07	**0.12**	0.06
Bioimaging and Data Visualization	0	8	10	10	**14**	0	**6.76**	4.3	3.37	4.11
Databases and Ontologies	0	**2**	0	0	**2**	0.07	**0.21**	0.05	0.13	**0.21**
Disease Models and Epidemiology	0	0	0	0	0	0	0	0	0	0
Evolution and Comparative Genomics	0	**2**	**2**	**2**	**2**	0	**1.14**	1.05	0.18	0.18
Gene Regulation and Transcriptomics	0	0	0	0	0	0	0	0	0	0
Mass Spectrometry and Proteomics	**6**	2	2	0	0	**0.75**	0.67	0.25	0.26	0
Metabolic Networks	4	10	12	14	**16**	3.52	13.56	**18.75**	17.59	14.27
Population Genomics	2	**4**	2	2	0	0.04	**0.38**	0.12	0.05	0
Protein Interactions and Molecular Networks	4	4	6	**12**	**12**	1.08	0.51	1.44	**4.89**	2.29
Protein Structure and Function	6	6	8	14	**16**	0.56	1.2	2.26	**3.83**	3.46
RNA Bioinformatics	0	0	8	**16**	14	0	0	2.89	**17.49**	16.59
Sequence Analysis	0	0	0	0	0	0.02	0	**0.04**	0.03	0
Text Mining	10	12	12	14	**16**	3.93	8.5	7.95	15.64	**16.77**
Average	2.57	3.86	4.71	6.14	**6.71**	0.72	2.36	2.8	**4.54**	4.14

"Population Genomics", none of the four methods found any expert in top 50 results. In the remaining 12 topics, BMExpert was the best performed method in seven topics, and both Jane and GoPubMed performed best in only three topics (with one tie). We obtained similar results in terms of AP. BMExpert achieved the highest AP of 4.14%, followed by Jane(1.96%), eTBLAST (1.73%) and GoPubMed (1.68%), respectively. Overall, we can see that, by modeling the expert finding problem with a formal language model, BMExpert outperformed all three existing biomedical expert finding services of Jane, eTBLAST and GoPubMed. The performance of BMExpert resulted from the way of computing three factors: 1) the importance of the document, 2) the association between the document and the expert, and 3) the relevance between the document and the topic query. In this paper, the relevance between the document and the topic query was computed by the language model. In the following, we explored the effects of the other two factors to the performance of BMExpert.

For incorporating the importance of the document, we varied several options: 1) recentness: the publication date of the document; 2) IF: the SCI impact factor of journal in which the document appears; and 3) domain: the document appearing in the journals in a specific domain.

Table 3 shows the results obtained by examining recentness. That is, we examined the performance of BMExpert by only considering the paper published in the last few years (1, 3, 5, 10 or all). We can see that only considering the paper published recently (in the last 1, 3 or 5 years) lowered the performance of BMExpert. Specifically, BMExpert with considering all papers achieved the highest P@50 of 6.71%, which was followed by considering papers published in the last ten years (6.14%), five years (4.71%), three years (3.86%) and one year (2.57%). Similar results were obtained when the performances were measured by AP. The only difference is that considering the papers published in the last ten years achieved the highest average AP of 4.54%, while considering all papers became the second best with the average AP of 4.14%. These results indicate that the true experts usually publish related papers constantly for a long term.

Table 4 shows the results by studying the effect of considering IF and domain (dm), i.e. a specific journal list in the bioinformatics domain. In this table, we compared the performance of various combinations of different settings: IF, no IF and no dm (note that this setting is the same as in Table 2), IF and dm, and dm. From this table, by restricting the papers in bioinformatics related journals, the

TABLE 4
The performances of BMExpert with considering impact factor (IF) and a specific journal domain (dm) (%).

Topics	P@50				AP			
	IF	no IF (and no dm)	IF and dm	dm	IF	no IF (and no dm)	IF and dm	dm
Applied Bioinformatics	2	2	2	**4**	0.15	0.06	**0.7**	0.4
Bioimaging and Data Visualization	10	14	**16**	12	4.16	4.11	**9.96**	7.86
Databases and Ontologies	**4**	2	**4**	2	0.56	0.21	**0.75**	0.54
Disease Models and Epidemiology	0	0	0	0	**0.03**	0	**0.03**	0
Evolution and Comparative Genomics	2	2	6	**8**	0.2	0.18	0.75	**0.87**
Gene Regulation and Transcriptomics	0	0	0	0	0	0	0	0
Mass Spectrometry and Proteomics	0	0	**6**	**6**	0.26	0	**0.75**	0.7
Metabolic Networks	**16**	**16**	14	**16**	**14.7**	14.27	13.89	13.82
Population Genomics	0	0		0	0	0	**0.09**	0
Protein Interactions and Molecular Networks	8	12	10	**14**	2.65	2.29	5.81	**6.08**
Protein Structure and Function	14	**16**	14	12	2.84	3.46	2.84	**6.27**
RNA Bioinformatics	**16**	14	10	10	**17.62**	16.59	10.56	10.89
Sequence Analysis	0	0	0	0	0	0	0	0
Text Mining	**16**	**16**	12	12	13.79	**16.77**	15.04	15.1
Average	6.29	6.71	6.71	**6.86**	4.07	4.14	4.37	**4.47**

TABLE 5
The performances of BMExpert with considering author positions (%).

Topics	P@50					AP				
	first	last	first and last	all equally	all eTBLAST	first	last	first and last	all equally	all eTBLAST
Applied Bioinformatics	0	2	2	**4**	0	0.02	**0.15**	0.06	0.1	0.01
Bioimaging and Data Visualization	2	10	**14**	6	10	0.25	**4.24**	4.11	1.86	3.82
Databases and Ontologies	2	0	2	**4**	2	0.23	0.06	0.21	**0.65**	0.23
Disease Models and Epidemiology	0	0	0	0	0	0	0	0	0	0
Evolution and Comparative Genomics	0	**2**	**2**	**2**	**2**	0.15	0.2	0.18	0.05	**0.32**
Gene Regulation and Transcriptomics	0	0	0	0	0	0	0	0	0	0
Mass Spectrometry and Proteomics	0	**2**	0	0	0	0.07	**0.24**	0	0.05	0.15
Metabolic Networks	6	14	**16**	14	**16**	3.28	11.34	14.27	11.59	**15.67**
Population Genomics	0	0	0	0	0	0	**0.02**	0	**0.02**	0
Protein Interactions and Molecular Networks	6	6	**12**	10	**12**	0.94	1.62	2.29	**2.47**	2.24
Protein Structure and Function	8	12	**16**	12	12	2.26	2.89	**3.46**	2.23	2.95
RNA Bioinformatics	6	14	14	14	**16**	5.92	18.12	16.59	18.89	**20.78**
Sequence Analysis	0	0	0	0	0	0	0	0	0	0
Text Mining	14	16	16	**18**	12	6.67	12.66	16.77	**20.46**	16.87
Average	3.14	5.57	**6.71**	6	5.86	1.41	3.68	4.14	4.17	**4.5**

performance of BMExpert was improved in both P@50 and AP. For example, with incorporating IF, BMExpert with considering all journals achieved an average P@50 of 6.29%, while BMExpert with considering only bioinformatics related journals achieved a higher average P@50 of 6.71%. Similarly, the average AP was improved from 4.07% to 4.37%. On the other hand, the effect of considering IF was mixed. Although the average AP was slightly decreased by incorporating IF, the AP of most topics were improved. For example, with considering all journals, the average AP of BMExpert was decreased from 4.14% to 4.07%. However, the APs of 9 out of all 14 topics (with 2 ties) were improved by incorporating IF. This suggests that incorporating IF tends to improve the AP for most topics.

For considering the associations between the document and the author, we examined five options: 1) all authors equally; 2) only the first author; 3) only the last author; 4) the first and last authors and 5) all authors with eTBLAST weighting scheme. Table 5 shows the performance results obtained under these five settings. From this table, considering both the first and last authors achieved the highest average P@50 of 6.71%, which was followed by considering all authors equally (P@50 of 6%), all authors with eTBLAST weighting scheme (P@50 of 5.86%), the last author only (P@50 of 5.57%) and the first author only (P@50 of 3.14%), respectively. On the other hand, considering all authors with eTBLAST weighting scheme achieved the highest average AP of 4.5%, which was followed by considering all authors equally (AP of 4.17%), both the first and last authors (AP of 4.14%), the last author only (AP of 3.68%) and the first author only (AP of 1.41%). From the experimental results we can see that both the first and last authors are the most important to the performance of BMExpert, and other authors also contribute to the performance, especially for AP.

Finally we examined the performance of BMExpert under different settings for improving P@50 and AP. According to the previous experiments, we chose the following settings in BMExpert to achieve better P@50: no published date restriction, considering bioinformatics related journals, no IF, and considering both the first and last authors, which we denote as BMExpert(P@50). On the other hand,

TABLE 6
Performance comparison of all competing methods in terms of P@50(%). BMExpert(P@50) and BMExpert(AP) are two new settings of BMExpert to achieve better P@50 and APo, respectively.

Topics	P@50					
	BMExpert	BMExpert(P@50)	BMExpert(AP)	Jane	GoPubMed	eTBLAST
Applied Bioinformatics	2	4	4	2	**6**	2
Bioimaging and Data Visualization	**14**	12	12	0	2	4
Databases and Ontologies	2	2	2	4	**8**	0
Disease Models and Epidemiology	0	0	0	0	0	0
Evolution and Comparative Genomics	2	**6**	**6**	0	0	0
Gene Regulation and Transcriptomics	0	2	0	**2**	0	0
Mass Spectrometry and Proteomics	0	**6**	4	2	2	2
Metabolic Networks	**16**	14	**16**	6	8	2
Population Genomics	0	**4**	2	0	0	0
Protein Interactions and Molecular Networks	12	12	**16**	4	10	8
Protein Structure and Function	**16**	12	10	6	2	0
RNA Bioinformatics	**14**	12	10	6	12	0
Sequence Analysis	0	**2**	0	**2**	0	0
Text Mining	**16**	12	12	10	4	12
Average	6.71	**7**	6.71	3.14	3.86	2.14

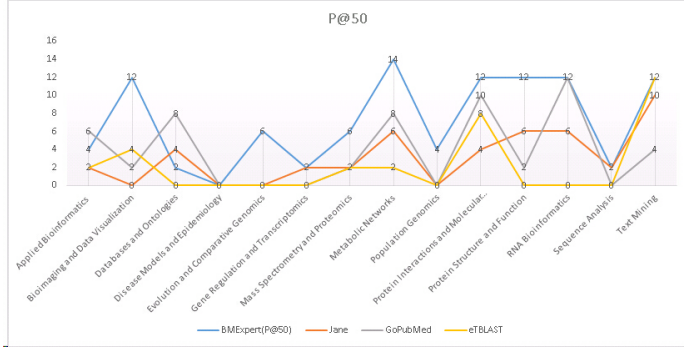

Fig. 1. Performance comparison of BMExpert(P@50) with Jane, eTBLAST and GoPubMed in terms of P@50.

we chose the following settings in BMExpert to achieve better AP: paper published in the last ten years, considering bioinformatics related journal, IF, and considering all authors with eTBLAST weighting scheme, which we denote as BMExpert(AP). Table shows the performance results of these two settings as well as BMExpert, Jane, GoPubMed and eTBLAST in terms of P@50. Table shows the performance comparison of BMExpert, BMExpert(P@50), BMExpert(AP), eTBLAST, Jane and GoPubMed in terms of P@50. Figure 1 shows visualization of the results (in Table) by BMExpert(P@50), Jane, GoPubMed and eTBLAST. We can see that BMExpert(P@50) achieved the highest average P@50 of 7%, which was followed by BMExpert (6.71%), BMExpert(AP) (6.71%), GoPubMed (3.86%), Jane (3.14%) and eTBLAST (2.14%). Similarly, we compared the performance of these methods in terms of AP. Table 7 shows the results of the performance comparison in terms of AP, and Figure 2 shows the visualization of the results in Table 7. From these table and figure, we can see that BMExpert(AP) achieved the highest average AP of 4.83%, which was followed by BMExpert(P@50) (4.74%), BMExpert (4.14%), Jane (1.96%), eTBLAST (1.73%) and GoPubMed (1.68%). Specifically, BMExpert(AP) outperformed BMExpert in 9 out of 14 topics (with 1 tie). These results suggest that the performance of BMExpert could be further improved with suitable configurations of three important factors.

Discussion and Conclusion

Our experimental results have demonstrated that BMExpert outperformed three existing biomedical expert finding services: Jane, GoPubMed and eTBLAST in both P@50 and AP. In contrast to three existing methods that use *ad hoc* approaches, BMExpert integrated three important factors simultaneously by a weighted language model, which achieved good performance in finding experts. These three factors include the importance of documents, the association between the experts (authors) and the document, and the relevance between the query and documents. In this paper, we focused on the effect of the first two factors. For considering the document importance, we examined three properties of documents, i.e. recentness, IF and the specific domain of documents. Since the topics in our benchmark dataset are from bioinformatics domain, it is not surprising that the performance of BMExpert was improved in both P@50 and AP, after we restricted the relevant papers appearing in bioinformatics related journals. Also we found that emphasizing most recently published papers is not very helpful for finding experts, which suggests that experts usually publish related papers constantly for a relatively long period. The effect of incorporating SCI impact factors was mixed. Although P@50 was slightly decreased, AP

TABLE 7
Performance comparison of all competing methods in terms of AP(%). BMExpert(P@50) and BMExpert(AP) are two new settings of BMExpert to achieve better P@50 and AP, respectively.

Topics	AP					
	BMExpert	BMExpert(P@50)	BMExpert(AP)	Jane	GoPubMed	eTBLAST
Applied Bioinformatics	0.06	0.72	0.06	0.06	**0.77**	0.01
Bioimaging and Data Visualization	4.11	**8.32**	8.31	0.05	0.12	0.5
Databases and Ontologies	0.21	0.67	0.96	1.02	**1.39**	0.04
Disease Models and Epidemiology	0	**0.03**	**0.03**	0	0	0
Evolution and Comparative Genomics	0.18	0.87	**1.01**	0	0	0
Gene Regulation and Transcriptomics	0	0	0	**0.04**	0	0
Mass Spectrometry and Proteomics	0	1.05	0.84	**1.33**	0.11	0.38
Metabolic Networks	14.27	14.16	**14.73**	2.52	3.11	0.65
Population Genomics	0	**0.42**	0.33	0	0	0
Protein Interactions and Molecular Networks	2.29	5.18	**6.96**	0.41	1.22	4.77
Protein Structure and Function	3.46	**7.62**	6.42	0.49	0.32	0
RNA Bioinformatics	**16.59**	12.06	13.75	8.03	13.53	0
Sequence Analysis	0	**0.2**	0	0.08	0	0
Text Mining	16.77	15.05	14.19	13.71	2.77	**17.85**
Average	4.14	4.74	**4.83**	1.96	1.68	1.73

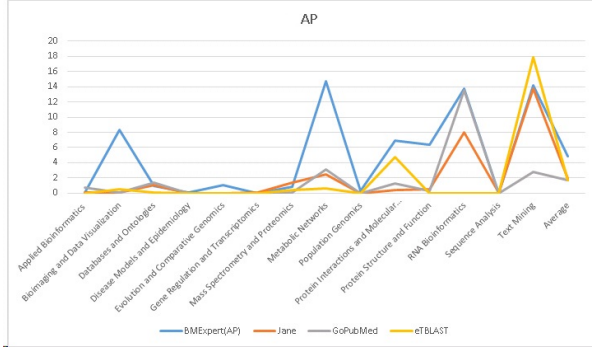

Fig. 2. Performance comparison of BMExpert(AP) with Jane, eTBLAST and GoPubMed in terms of AP.

was increased for most of the topics. This indicates that many experts publish papers in high profile journals. For measuring the associations between experts and documents, we found that the first and last authors are the most important expert candidates by examining various options. This is intuitive in biomedical domain, since the first and last authors are usually most familiar with the paper. The first author carries out experiments and the last author is the supervisor who proposed and leaded the project and gave the guidance. Interestingly, only considering either of them misses some important information.

The relevance between a document and a query topic in BMExpert was computed based on the language model. According to the generalized document based framework for expert finding (GDFEF), we may also resort to other methods for computing the relevance between a document and a query topic. In fact, eTBLAST uses sentence alignment scoring and Jane uses Lucene scoring for computing the relevance between a document and a query. It would be interesting to examine the performance of BMExpert by replacing the language model with some other standard information retrieval models, such as vector space model and Okapi BM25 model [18]. Compared with existing methods, BMExpert tends to perform very well on specific topics, such as "RNA Bioinformatics", "Metabolic Networks" and "Text Mining". On the other hand, there are some topics on which none of the four competing methods performed well, such as "Disease Models and Epidemiology", "Evolution and Comparative Genomics" and "Sequence analysis". This result might be caused mainly by the following three reasons. Firstly, there are some phrases in the query topic, while existing methods cannot handle them efficiently. Secondly, the query terms do not appear in the document, while their synonyms appear. In this case, query expansion techniques should be used to retrieve the relevant documents. Finally, understanding the user's intension from the input query topic is not a trivial task. In this case, advanced techniques in information retrieval are needed to reformulate the input query for further processing.

In this work we have introduced BMExpert, a novel solution for finding experts in biomedical domain through mining MEDLINE. Biomedical science is a wide, fast growing field, which results in that finding biomedical experts is getting more important. BMExpert is based on a weighted language model, which can consider author weighting, paper importance and paper relevance. Our evaluation experiments have shown that BMExpert clearly outperformed the three other competing methods under the gold-standard dataset derived from programming committee members of the ISMB conferences which were held in the past three years. We believe that BMExpert must be useful

for biomedical researchers who want to find experts in biomedical domain. Important future work would be to use more complex queries than current phrase-oriented queries. To realize complex queries, a possible way would be to consider logic relations to generate queries.

Acknowledgement

This work has been partially supported by National Natural Science Foundation of China (61170097), and Scientific Research Starting Foundation for Returned Overseas Chinese Scholars, Ministry of Education, China. H.M. is partially supported by JSPS KAKENHI (#2430054). S.Z. would like to thank the China Scholarship Council for the financial support on his visit at University of Illinois at Urbana-Champaign. We would like to thank Fudan University Library for providing us the SCI Impact Factor for academic usage.

References

[1] E. Sayers, T. Barrett, D. Benson, E. Bolton, S. Bryant, K. Canese, V. Chetvernin, D. Church, M. DiCuccio, S. Federhen, and others., "Database resources of the national center for biotechnology information." *Nucleic Acids Research*, vol. 39, pp. D38–D51, 2011.

[2] A. Eaton, "HubMed: a web-based biomedical literature search interface." *Nucleic Acids Research*, vol. 34, pp. W745–W747, 2006.

[3] A. Doms and M. Schroeder, "GoPubMed: Exploring PubMed with the gene ontology." *Nucleic Acids Research*, vol. 33, pp. W783–W786, 2005.

[4] C. Perez-Iratxeta, A. Perez, P. Bork, and M. Andrade, "Update on XplorMed: a web server for exploring scientific literature." *Nucleic Acids Research*, vol. 31(13), pp. 3866–3868, 2003.

[5] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu, "Efficient semi-supervised MEDLINE document clustering with MeSH semantic and global content constraints," *IEEE Transactions on Cybernetics*, vol. 43 (4), pp. 1265–1276, 2013.

[6] S. Zhu, J. Zeng, and H. Mamitsuka, "Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity," *Bioinformatics*, vol. 25(15), pp. 1944–1951, 2009.

[7] S. Zhu, I. Takigawa, J. Zeng, and H. Mamitsuka, "Field independent probabilistic model for clustering multi-field documents," *Information Processing & Management*, vol. 45(5), pp. 555–570, 2009.

[8] X. Huang, X. Zheng, W. Yuan, F. Wang, and S. Zhu, "Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization," *Information Science*, vol. 181(11), pp. 2293–2302, 2011.

[9] I. Iliopoulos, A. Enright, and C. Ouzounis, "Textquest: document clustering of Medline abstracts for concept discovery in molecular biology," *Pac. Symp. Biocomput*, vol. 3, pp. 84–95, 2001.

[10] K. Liu, J. Wu, S. Peng, C. Zhai, and S. Zhu, "The Fudan-UIUC participation in the BioASQ Challenge Task 2a: The antinomyra system," *CLEF (Working Notes)*, pp. 1311–1318, 2014.

[11] B. Krisztian, F. Yi, d. R. Maarten, S. Pavel, and S. Luo, "Expertise retrieval," *Foundations and Trends in Information Retrieval*, vol. 6, pp. 127–256, 2012.

[12] M. Errami, J. Wren, J. Hicks, and H. Garner, "eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications." *Nucleic Acids Research*, vol. 35, pp. W12–W15, 2007.

[13] M. Schuemie and J. Kors, "Jane: suggesting journals, finding experts." *Bioinformatics*, vol. 24(5), pp. 727–728, 2008.

[14] E. Hatcher, O. Gospodnetic, and M. McCandless, *Lucene in action*. 209 Bruce Park Avenue, Greenwich, CT 06830: Manning Publications, 2004.

[15] M. Plikus, Z. Zhang, and C.-M. Chuong, "PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm." *BMC Bioinformatics*, vol. 7(1), p. 424, 2006.

[16] K. Balog, L. Azzopardi, and M. de Rijke., "A language modeling framework for expert finding." *Information Processing and Management*, vol. 45, pp. 1–19, 2009.

[17] H. Deng, I. King, and M. Lyu, "Formal models for expert finding on DBLP bibliography data," *ICDM'08*, pp. 163–172, 2008.

[18] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Corn Exchange Street: Cambridge University Press, 2008.

Beichen Wang received the B.S. degree in computer science from the Fudan University, Shanghai, China, in 2012. He is currently working toward the M.Eng. degree in Shanghai Key Laboratory of Intelligent Information Processing and School of Computer Science, Fudan University, Shanghai, China. His research interests are machine learning, information retrieval, and bioinformatics.

Xiaodong Chen received the B.S. degree in Medicine and M.Eng. degree in computer sciece from Fudan University, Shanghai, China, in 2010 and 2013 respectively. His research interests are machine learning, information retrieval, and bioinformatics.

Hiroshi Mamitsuka received the B.S. degree in biophysics and biochemistry, the M.E. degree in information engineering, and the Ph.D. degree in information sciences from the University of Tokyo, Tokyo, Japan, in 1988, 1991, and 1999, respectively. He is currently a Professor with Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. His current research interests include mining from graphs and networks in biology and chemistry.

Shanfeng Zhu Shanfeng Zhu received the B.S. and M.Phil. degrees in computer science from Wuhan University,Wuhan, China, in 1996 and 1999, respectively, and the Ph.D. in computer science from the City University of Hong Kong, Hong Kong, in 2003. He was a Postdoctoral Fellow with the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. Since July 2008, he has been with the School of Computer Science and Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China, where he is currently an Associate Professor. His research focuses on developing and applying machine learning, data mining, and algorithmic methods for bioinformatics and information retrieval. Dr. Zhu is a member of China Computer Federation and the Association for Computing Machinery.