Finding Topic-level Experts in Scholarly Networks

Lili Lin[1,*], Zhuoming Xu[1], Ying Ding[2], Xiaozhong Liu[2]

[1]College of Computer and Information, Hohai University, Nanjing, China

[2]School of Library and Information Science, Indiana University, Bloomington, USA

*Corresponding author: linlili@hhu.edu.cn, College of Computer and Information, Hohai University, No. 8 Focheng West Road, Jiangning District, Nanjing 211100, Jiangsu Province, China

Abstract

Expert finding is of vital importance for exploring scientific collaborations to increase productivity by sharing and transferring knowledge within and across different research areas. Expert finding methods, including content-based methods, link structure-based methods, and a combination of content-based and link structure-based methods, have been studied in recent years. However, most state-of-the-art expert finding approaches have usually studied candidates' personal information (e.g. topic relevance and citation counts) and network information (e.g. citation relationship) separately, causing some potential experts to be ignored. In this paper, we propose a Topical and Weighted Factor Graph (TWFG) model that simultaneously combines all the possible information in a unified way. In addition, we also design the Loopy Max-Product algorithm and related message-passing schedules to perform approximate inference on our cycle-containing factor graph model. Information Retrieval is chosen as the test field to identify representative authors for different topics within this area. Finally, we compare our approach with three baseline methods in terms of topic sensitivity, coverage rate of SIGIR PC (e.g. Program Committees or Program Chairs) members, and NDCG (Normalized Discounted Cumulated Gain) scores for different rankings on each topic. The experimental results demonstrate that our factor graph-based model can definitely enhance the expert-finding performance.

Keywords: expert finding, factor graph, topic relevance, scholarly network

Introduction

In order to make good use of expertise and knowledge, an important task in scientific research area named *expert finding* or *expert searching* has received a significant amount of attention in recent years. The goal of expert finding is to return a ranked list of knowledgeable experts with relevant expertise on a specific topic or research area. This expert finding process can help solve many challenging but practical problems. For example, in order to improve the quality of published papers and to facilitate dissemination of accurate and valid knowledge to a research area/topic, peer review process has long been strongly recommended. One of the most important elements of this process is how to model the expertise of a given reviewer with respect to the topical content of a given paper. However, matching papers with suitably qualified reviewers is still a challenging process (Mimno & McCallum, 2007). Other important applications include recommending the reviewers for the evaluation of research grant applications (Hettich & Pazzani, 2006), determining important experts for consultation by researchers embarking on a new research field (Serdyukov, Henning & Hiemstra, 2008), recruiting employees for one certain job position etc. However, manually identifying these experts in a large research area or organization is obviously labor intensive and time consuming. A standard text search engine may be of great help, but it is still not able to automate this task (Serdyukov, Henning & Hiemstra, 2008). It is therefore meaningful and even essential to study and ascertain how to automatically identify experts on a specific topic on a large scale.

Some researchers use content-based methods to detect persons who are experts on a specific topic. However, these kinds of methods mostly concentrate on providing relevance scores between candidates and a user's query topic or an inferred topic, while neglecting the social relationships between candidates for more precise expert identification. Another option is to use link analysis algorithms such as PageRank (Page, Brin, Motwani & Winograd, 1999) and HITS (Kleinberg, 1999) to address expert-finding tasks. But PageRank and HITS have a common problem: topic drift, which tends to make most in-links in the network dominant (Zhang, Tang & Li, 2007). Due to the limitations of content-based methods and traditional link structure-based methods, some previous works (Campbell, Maglio, Cozzi & Dom, 2003; Zhang, Tang & Li, 2007; Jiao, Yan, Zhao & Fan, 2009; Tang, Sun, Wang & Yang, 2009; Ding, 2011) not only consider the relevance of a candidate on a specific topic, but also analyze scholarly networks between candidates in order to improve expert finding efficiency. To the best of our knowledge, however, most of these methods model possible information separately and then combine it in a specific sequence, which causes possible experts to be ignored.

Motived by observations on certain common characteristics of judgments people make to find experts, we define two important features, topic relevance and expert authority, as personal information and extract the citation relationships between authors to build the citation network. As factor graphs have the potential to unify modeling with great generality and flexibility (Kschischang, Frey & Loeliger, 2001), we propose a Topical and Weighted Factor Graph (TWFG) model to tackle these expert-finding limitations, and also design the Loopy Max-Product algorithm with serial schedule using random sequences to perform approximate inference on our cycle-containing factor graph model. Our approach is unique in that it correlates all candidates' personal and network information into a unified model based on factor graph theory (Kschischang, Frey & Loeliger, 2001; Bishop, 2006) and conducts inference in a global manner. Another distinguishing feature of our proposed approach is its modeling of mutual influences between candidates on a topic level, which is quite different from current approaches.

The remainder of this paper is organized as follows: Section 2 reviews related work on various expert-finding methods. Section 3 explicates the proposed approach. Section 4 presents the experimental results that validate the efficiency of our methodology. Finally, we conclude our work in Section 5.

Related Work

Several studies have investigated approaches for expert finding. The existing approaches can be divided into three main categories according to their focuses. The comparison of existing expert finding approaches is shown as Table 1.

Content-based Methods

Much attention has been given to content-based models for expert finding. Some models typically fall into one of the two classifications–those that generate the probability of a candidate being an expert given a user's query, and those that generate this probability based on the latent topic variables inferred from word correlations.

The first kind of content-based methodology is treated as an information retrieval task by Text REtrieval Conference (TREC). Such methods are basically variations of two kinds: profile-centric methods (also referred to as candidate-centric or query-independent approaches) and document-centric methods (also referred to as query-dependent approaches) (Petkova & Croft, 2006; Balog, Azzopardi & Rijke, 2009; Smirnova & Balog, 2011). In profile-centric methodologies (Balog, Azzopardi & Rijke, 2006; Fu, Xiang, Liu, Zhang & Ma, 2007), all documents or texts related to a candidate are first merged into a single personal profile, where the ranking score for each candidate is then estimated according to the profile in response to a given query. Nevertheless, these document-centric methods (Balog, Azzopardi & Rijke, 2006; Wu, Pei & Yu, 2009) analyze the content of each document separately instead of creating a single expertise profile. In order to make use of the advantages of both the profile-centric and document-centric methods, some existing approaches (Petkova & Croft, 2006; Serdyukov, Henning & Hiemstra, 2008) combine the two methods to improve expert-finding performance. However, these kinds of studies generally concentrate on aligning search results with user queries, which are different from topic-dependent expert finding based on automatically inferred latent topics.

The second kind of content-based methods is known as topic modeling. An early topic model, named Probabilistic Latent Semantic Indexing (PLSI), was proposed by Hofmann (1999) to calculate the probability of generating a word from a document based on the latent topic layer. However, the parameterization of the PLSI model is susceptible to severe overfitting, and the PLSI model does not provide a straightforward way to make inferences about documents. Blei, Ng, and Jordan (2003) addressed these limitations by proposing a three-level hierarchical Bayesian model called latent Dirichlet allocation (LDA). As a follow-up effort of the LDA model, Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) introduced the author-topic model to depict the content of documents and the interests of authors simultaneously by sharing the hyperparameters of topic mixing for all documents by the same authors. Specifically, each author is associated with a multinomial distribution over topics, allowing the clusters of authors to be detected. Later, Tang, Jin and Zhang (2008) further extended the LDA and author-topic model and proposed the Author-Conference-Topic (ACT) model to organize different types of information concurrently in academic networks. The ACT model includes publication venues, so that each author is associated with a multinomial distribution over topics, words he/she wrote, and conferences in which he/she was published. Not surprisingly, topic models can help calculate the relevance between candidates and an inferred topic for further ranking. However, they don't investigate the relationships between the knowledge presentations of candidates to build up a scholarly network for more sophisticated expert-evidence identification and extraction.

Link Structure-based Methods

As content related to candidates cannot serve as direct evidence of their expertise, a few studies have tried to employ link structure among candidates to address the expert-finding problem. Link structure-based algorithms, such as PageRank (Page, Brin, Motwani & Winograd, 1999) and HITS (Kleinberg, 1999), can be used to analyze relationships in a scholarly network in order to find authorized experts. Liu, Bollen, Nelson and Sompel (2005) developed AuthorRank for this purpose, a modification of PageRank that considers link weights among the coauthorship links. Based on traditional link analysis-ranking algorithms, Sidiropoulos and Manolopoulos (2006) developed a new method specifically designed for citation graphs to evaluate the impact of scientific collections (journal and conferences), publications, and scholarly authors. They also introduced an aggregate function for the generation of author ranking based on publication ranking. Jurczyk and Agichtein (2007) explored the HITS link analysis algorithm to estimate the authority of users that can be potentially used for finding experts in Question Answer portals. Fiala, Rousselot, and Ježek (2008) presented several modifications of the classical PageRank formula adapted for bibliographic networks. Their ranking results based on both on the citation and co-authorship information turned out to be better than the standard PageRank ranking. Ding, Yan, Frazho, and Caverlee (2010) used the PageRank algorithm with different damping factors and also proposed two different weighted PageRank algorithms to rank authors on an author co-citation network. A weighted PageRank algorithm that considers citation and coauthorship network topology was proposed by Yan and Ding (2011) to measure author impact. All those works aimed at applying variations of HITS or PageRank algorithms or other link-based methods in the context of author ranking in order to alleviate the limitations of some classical indicators (e.g. citation counts) for ranking in bibliometrics. However, they are not effective for finding the top "experts" without considering content features. Moreover, all of them are topic-independent, and include certain classical indicators such as impact factor, H-index, and citation counts.

Combination of Content-based and Link Structure-based Methods

Some researchers have used documents or snippet-level content to provide topic relevance for each candidate, and then applied link analysis to further refine the ranking results. Campbell, Maglio, Cozzi, and Dom (2003) used text analysis and network analysis to sort individuals within an email network. Specifically, they collected all emails related to a topic and analyzed emails between every pair of people for whom there was relevant correspondence to build an "expertise graph." They finally applied a modified HITS algorithm to obtain ratings for all senders and recipients on that topic. Zhang, Tang, and Li (2007) first used candidates' personal information (e.g. personal profile, contact information, and publications) to estimate an initial expert score for each candidate and selected the top ranked candidates to construct a subgraph. They then proposed a propagation-based approach to improve the accuracy of expert finding within the subgraph. Jiao, Yan, Zhao, and Fan (2009) used expert relevance scores to generate a subset of experts, and also used a modified PageRank algorithm to calculate the authority scores of experts. They then combined expert relevance and expert authority with a linear formula to express the final expertise of a candidate. Ding (2011) proposed topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm. Two ways for combining the ACT model with the PageRank algorithm are proposed in her work: simple combination or using a topic distribution as a weighted vector for PageRank. However, most of the above methods do not simultaneously model all the possible information in a unified way. Furthermore, all of them have been achieved by encountering local optimization by ranking within a limited subset of candidates. A notable exception is that of Tang, Sun, Wang, and Yang (2009), who proposed a topical factor graph model to identify representative nodes from scholarly networks on a specific topic by leveraging the topic relevance and social relationships between links. Yet because not all coauthored papers are highly correlated with a specific topic, we argue that it is not reasonable to weigh edges between candidates mainly on coauthored papers to reflect the interaction strength between neighboring nodes. Moreover, this research tends to use a subset of candidates for identifying representative authors, which may filter out some potential experts.

Table 1 Comparison of existing expert finding approaches

Category	**Work**	**Topic-dependent**	**Social Network**	**Model or Algorithm**	**Other**
Content-based Methods	Balog, et al, 2006	**Yes** (assign users' queries as topics)	None	**Profile-centric model**: constructs an expertise profile; assesses how probable the query topic is to rank candidates **Document-centric model**: ranks documents according to user query; ranks candidates by considering their associated documents	No link analysis for more sophisticated expert-evidence identification and extraction
	Fu, et al, 2007			Profile-centric model	
	Wu, et al, 2009			Document-centric model	
	Petkova, et al, 2006			Combination of the profile-centric and document-centric model	
	Rosen-Zvi, et al, 2004	**Yes** (automatically infer latent topics)		**Author-Topic Model:** simultaneously depict the content of documents and the interests of authors	
	Tang, et al, 2008			**Author-Conference-Topic (ACT) Model:** simultaneously model papers, authors, and paper venues	
Link Structure-based Methods	Liu, et al, 2005	**No**	Coauthor network	AuthorRank algorithm by modifying PageRank algorithm	Tend to make most in-links in the network dominant
	Sidiropoulos, et al, 2006		Citation network	Balanced HITS algorithm	
	Jurczyk, et al, 2007		Question-answer user network	HITS algorithm	
	Fiala, et al, 2008		Coauthor network	Several modifications of PageRank algorithms	

	Ding, et al, 2010		Author co-citation network	Two weighted PageRank algorithms	
	Yan, et al, 2011		Coauthor network	A weighted PageRank algorithm	
Combination of Content-based and Link Structure-based Methods	Campbell, et al, 2003	**Yes** (topics are given by user)	Email network	A modified HITS algorithm	Ranking within a subset of candidates may filter out some potential experts
	Zhang, et al, 2007		Coauthor network	A propagation-based approach on a candidate subgraph.	
	Jiao, et al, 2009		Online communities	A modified PageRank algorithm on a candidate subgraph	
	Ding, 2011	**Yes** (topics are inferred by ACT model)	Author co-citation network	Two algorithms by combination of the ACT model and the PageRank algorithm on an highly cited author subgraph	
	Tang, et al, 2009		Author citation/ paper citation network	A topical affinity prorogation method on a candidate subgraph	

Methodology

In this section, we detail the ways in which we solve the expert-finding problem. First, we describe our data collection and present the problem formulation to define the expert-finding task. Second, three main modules of the proposed approach are presented in detail, including topic distribution, the topical and weighted factor graph (TWFG) model, and the inference mechanism.

Data Collection

In this paper, we choose Information Retrieval (IR) as the test field. Papers and their citations were collected from the Web of Science (WOS) covering the period from 2001 to 2008, including 8,396 papers and 14,593 authors with 211,560 citations among authors. Each paper contains related authors, title, source, published year, abstract, reference, citation counts, and so forth. The titles are preprocessed using a stemming algorithm and a stop word list. Citation records include the first author, published year, source, volume, and page number. Citations are used to generate a citation network. Details of our data collection are provided in Ding and Cronin (2010). In order to make our approach easier to describe and understand, the notations are first given in Table 2.

Table 2 Notations

Symbol	Description
N	the number of authors in the citation network
V	the set of authors in the citation network

E	the set of edges in the citation network
Y	the set of hidden vectors for all author nodes
z	a research topic within a research area
t	the number of topics
v_i	an author node in the citation network
$\mathbf{y}_i$	the hidden vector for all topics on a given author v_i
$\mathrm{y}_{\mathrm{i}}^{\mathrm{z}}$	author v_i's importance weight on a given topic z
n_i	the ranking order of a given author v_i based on his or her citation counts
α_{iz}	the probability of an author v_i on a given topic z
e_{ij}	an edge between author v_i and author v_j
θ_{ij}^{z}	the dissimilarity weight associated with edge e_{ij} on a given topic z

Problem Formulation

It was pointed out by Fu, Xiang, Liu, Zhang, and Ma (2007) that in most cases expertise data is not fully documented, and usually only parts of a document are related to the expert whom it mentions. In order to better describe the knowledge of a particular researcher, the expertise representation should be established from multiple aspects. To the best of our knowledge, citation counts were firstly proposed by Gross and Gross (1927) to evaluate the importance of researchers' work and then became a widely used indicator for scientific impact. However, some researchers interested in measuring scientific impact doubted that citation counts can reflect the impact of scientific activity (Bornmann & Daniel, 2008) Based on some empirical findings, Kochen (1978) suggested the use of citation counts in combination with content analysis to modify the use of citation counts in research measurement. Moreover, a citation implies a relationship between a part or the whole of the cited document and a part or the whole of the citing document (Smith, 1981). The analysis of this kind of relationship is often used as a tool for evaluating the performance and measuring the impact of scientists, institutions, journals, regions etc (Matutinovic, 2007). Even though there are many reasons why citation relationship exists, the most obvious reason is that the citing document is highly relevant with the cited document. Therefore, it is critical to combine the citation counts, content analysis and citation/link analysis for measuring the researchers' impact. Based on above investigations, here we adopt two critical features as local/personal evidence to represent the authors' fundamental expertise, and simultaneously use the citation relationships between the authors as additional evidence to weigh their expertise on a given topic. The above local and network information can be further defined as follows:

Topic relevance. This local feature can be used to model the relevance between an author and a specific topic. Given a topic, the amount of information that an author's publications contain contributes to the presentation of how much of the required knowledge that author has. We assume that if an author possesses a higher probability α_{iz} on a given topic z, he/she is more likely to be an expert on topic z. Formally, each author $v_i \in V$ is associated with a t-dimensional topic distribution $\{\alpha_{iz}\}_{1 \le z \le t}$ where $\sum_{1 \le z \le t} \alpha_{iz} = 1$.

Expert authority. This local feature can be used to model the knowledge of an author. We make an assumption, in that among those authors with the same relevance on a given topic, the author with higher citation counts is more likely to be an expert since he/she tends to be a popular author in scientific research areas (Fu, Xiang, Liu, Zhang & Ma, 2007; Ding & Cronin, 2010). It should be noted that here we use the ranking order n_i of each author based on citation counts to represent his/her expert authority.

Topic-level influences. Even with the same citation network structure, mutual influences between authors will vary on different topics. More precisely, when calculating author expertise on different topics, dissimilarities or similarities between authors can result in different contributions. Here, each edge e_{ij} on a given topic z can be denoted as $(v_i, v_j, \theta_{ij}^z)$, where the edge e_{ij} between author v_i and author v_j with dissimilarity weight θ_{ij}^z.

Based on the above definitions, we jointly take into account all the personal information and network information to formulate the problem as follows:

Given (1) a citation network $G = (V, E)$ where $V = \{v_i\}_{i=1}^N$ is the set of authors and $E = \{e_{ij}\}_{1 \le i,j \le N}$ is the set of edges representing citation relationships between authors, (2) t-dimensional topic distribution $\{\alpha_{iz}\}_{1 \le z \le t}$ for each author $v_i \in V$, and (3) the ranking order n_i of each author $v_i \in V$. The goal is to find topic-level ranked experts for each topic z $(1 \le z \le t)$ with the given personal information and citation network.

Topic Distribution

Latent Dirichlet allocation captures the topical features of nodes by postulating a latent structure for a set of topics linking words and documents (Blei, Ng &Jordan, 2003). As an extended LDA model, the Author-Conference-Topic model proposed by Tang, Jin, and Zhang (2008) can be used to capture more topical features of nodes. After applying the ACT model, five topics are automatically extracted. Simultaneously, each author’s topic distribution (i.e. the probability of an author writing on a given topic), the paper’s topic distribution (i.e. the probability of a paper being written on a given topic), and conference topic distribution (i.e. the probability of a conference taking place for a given topic) for all the extracted topics are calculated. Furthermore, each topic is associated with a list of words and a set of authors ranked by their topic distribution probabilities. In a citation network, an author often has interests on multiple topics. Here the ACT model can definitely help calculate the relevance between an author and related topics. For example, author A has a five-dimensional topic distribution {0.05, 0.6, 0.1, 0.07, and 0.18}. Among the extracted five topics, author A shows higher interests in topic 2 because the value of his/her probability on topic 2 is much higher than that of other topics. For author A, the sum of topic distribution across the five topics is 1.0 and the average probability for author A for these five topics would be 1/5=0.2.

Topical and Weighted Factor Graph Model

In order to combine all the possible information encoded in the citation network, we develop a topical and weighted factor graph model to leverage topic relevance, expert authority, and topic-level influence. For simplicity, we make an assumption that topics are independent of each other. Hence we can decompose our factor graph model into a set of factor graphs with the same topological structure on different topics. Fig.1 shows a simple TWFG on a given topic z corresponding to the example we have been used.

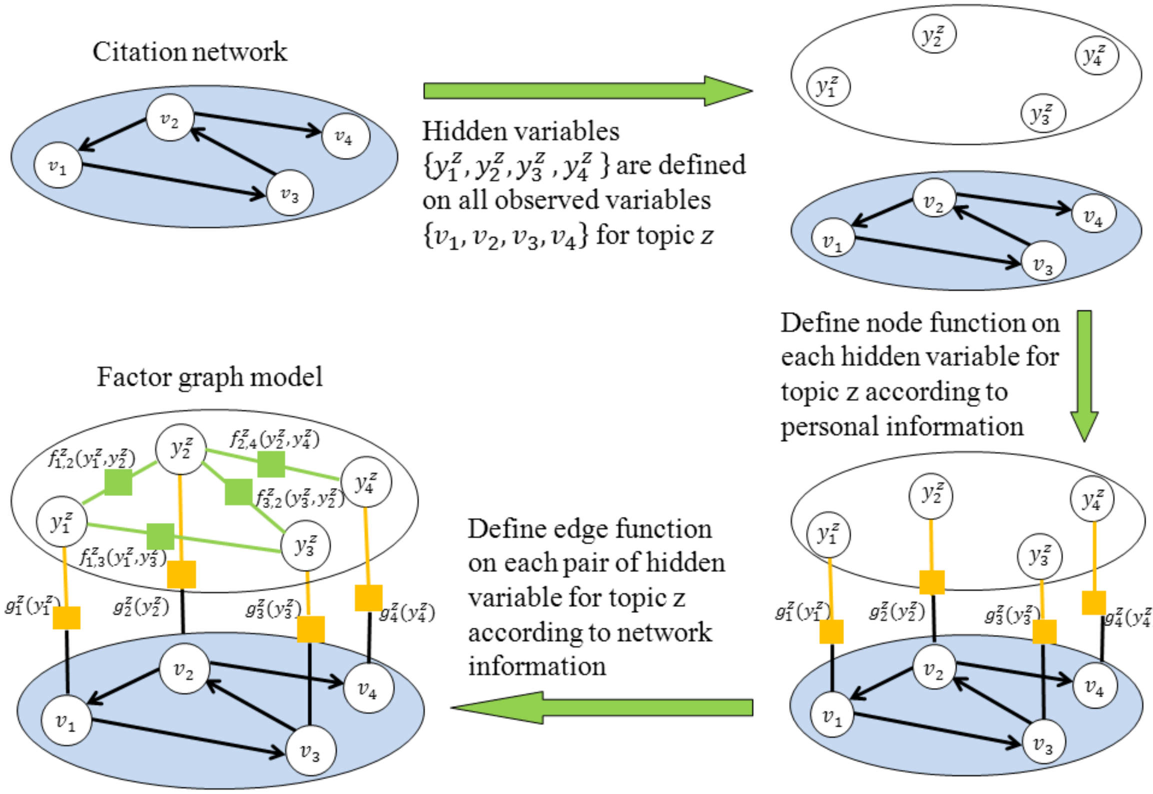
Citation network
v_1
v_2
v_3
v_4
Hidden variables $\{y_1^z, y_2^z, y_3^z, y_4^z\}$ are defined on all observed variables $\{v_1, v_2, v_3, v_4\}$ for topic z
y_1^z
y_2^z
y_3^z
y_4^z
Define node function on each hidden variable for topic z according to personal information
$g_1^z(y_1^z)$
$g_2^z(y_2^z)$
$g_3^z(y_3^z)$
$g_4^z(y_4^z)$
Define edge function on each pair of hidden variable for topic z according to network information
Factor graph model
$f_{1,2}^z(y_1^z, y_2^z)$
$f_{2,4}^z(y_2^z, y_4^z)$
$f_{3,2}^z(y_3^z, y_2^z)$
$f_{1,3}^z(y_1^z, y_3^z)$

Fig.1 Graphical representation of a topical and weighted factor graph on a given topic z, where $\{y_1^z, y_2^z, y_3^z, y_4^z\}$ are hidden variables defined on all observed variables $\{v_1, v_2, v_3, v_4\}$ for topic z; $g_i^z(.)$ represents a node function and $f_{ij}^z(.)$ represents an edge function

As each observed variable $v_i \in V$ corresponds to a hidden vector $\mathbf{y}_i \in Y$, the factor graph can be regarded as the composition of a set of hidden variables $Y = \{\mathbf{y}_i\}_{i=1}^N$ and a set of functions. Concretely, the functions in our model fall into node function g and edge function f. The former is used to model the personal information (i.e., topic relevance and expert authority) and the latter is used to model the network information (i.e., topic-level influences). Here, we define the **node function** as equation (1) on the intuition that authors with higher topic relevance on a given topic z are more likely to be experts on that topic and authors with higher citation counts tend to be experts even they have the same topic relevance:

$$g_i(\mathbf{y}_i, z) = g_i^z(y_i^z) = \begin{cases} \exp(n_i \alpha_{iz} y_i^z), \alpha_{iz} \geq \lambda \\ \exp(-n_i \alpha_{iz} y_i^z), \alpha_{iz} < \lambda \end{cases} \quad (1)$$

where n_i represents the ranking order of a given author v_i based on his/her citation counts; α_{iz} denotes the probability of an author v_i on a given topic z; $y_i^z \in \{0,1\}$ reflects the importance of an author for topic z, $y_i^z = 0$ indicates author v_i is not important for topic z and $y_i^z = 1$ indicates author v_i is important for topic z; and λ specifies the relevance threshold between an author and a topic, $\alpha_{iz} \geq \lambda$ indicates author v_i is more relevant with a given topic z and $\alpha_{iz} < \lambda$ indicates author v_i is less relevant with a given topic z.

Obviously, an important author v_i on a given topic z may have a high influence on one of his/her neighboring author node v_j if they have a high similarity on their research interests/topics. Author v_j may also then have a high probability to become an important author on topic z. In order to capture the topic-level influences between neighboring author nodes, we define **edge function** as equation (2).

$$f_{ij}(\,\mathbf{y}_i, \mathbf{y}_j, z) = f_{ij}^z(y_i^z, y_j^z) = \begin{cases} \exp(\theta_{ij}^z)\,, & \text{if } \theta_{ij}^z \leq \theta \text{ and } y_i^z = y_j^z \\ \exp(-\theta_{ij}^z), & \text{if } \theta_{ij}^z > \theta \text{ and } y_i^z = y_j^z \\ 1, & \text{if } y_i^z \neq y_j^z \end{cases} \qquad (2)$$

where $y_i^z \in \{0,1\}$ and $y_j^z \in \{0,1\}$ represents the importance weight of author v_i and author v_j on a given topic z, respectively; θ_{ij}^z indicates the dissimilarity weight between author v_i and author v_j on topic z, which is calculated based on K-L divergence (Kullback *et al.*, 1987) shown in equation (3); and θ specifies the dissimilarity threshold between author v_i and author v_j, $\theta_{ij}^z \leq \theta$ indicates author v_i and author v_j have more similar research interests on topic z and $\theta_{ij}^z > \theta$ indicates less similar research interests:

$$\theta_{ij}^z = \alpha_{iz} \ln \frac{\alpha_{iz}}{\alpha_{jz}} + \alpha_{jz} \ln \frac{\alpha_{jz}}{\alpha_{iz}} \qquad (3)$$

Based on above, we finally define the **objective function** by considering all the functions according to factor graph theory (Kschischang, Frey & Loeliger, 2001; Bishop, 2006) as seen in equation (4):

$$p(Y) = \frac{1}{S} \prod_{z=1}^{t} \prod_{i=1}^{N} g_i(\mathbf{y}_i, z) \prod_{z=1}^{t} \prod_{e_{ij} \in E} f_{ij}(\mathbf{y}_i, \mathbf{y}_j, z)$$

$$= \frac{1}{S} \prod_{z=1}^{t} \prod_{i=1}^{N} g_i^z(y_i^z) \prod_{z=1}^{t} \prod_{e_{ij} \in E} f_{ij}^z(y_i^z, y_j^z) \;= \frac{1}{S} \prod_{z=1}^{t} (\prod_{i=1}^{N} g_i^z(y_i^z) \prod_{e_{ij} \in E} f_{ij}^z(y_i^z, y_j^z)\;) \qquad (4)$$

where $Y = \{\mathbf{y}_1, \mathbf{y}_2 \cdots \mathbf{y}_N\}$ corresponds to all hidden variables; $g_i^z(y_i^z)$ is the node function and $f_{ij}^z(y_i^z, y_j^z)$ is the edge function; and S is a normalizing factor. As we have assumed that topics are independent, so that

$$p(Y) = \prod_{z=1}^{t} p(Y_z) \qquad (5)$$

Thus, once the topic is specified, the **objective function** for topic z can be defined as equation (6):

$$p(Y_z) = \frac{1}{S'} \prod_{i=1}^{N} g_i^z(y_i^z) \prod_{e_{ij} \in E} f_{ij}^z\left(y_i^z, y_j^z\right) \qquad (6)$$

where $Y_z = \{y_1^z, y_2^z, \cdots y_N^z\}$ corresponds to the hidden variables for topic z; and S' is a normalizing factor.

Inference Algorithm

As a generic message-passing algorithm, the Sum-Product algorithm (Kschischang, Frey & Loeliger, 2001) has often been applied to compute the marginals of all variable nodes efficiently and exactly for the factor graph-based model. The algorithm involves passing messages between variable nodes (i.e., hidden variables) and function nodes on the built factor graph (Kschischang, Frey & Loeliger, 2001). Message passing is initiated at the leaves. Each node v remains idle until messages have arrived on all but one of the edges incident on v. Once these messages have arrived, v is able to compute a message to be sent on the one remaining edge to its neighbor w (temporarily regarded as the parent). After sending a message to w, node v returns to the idle state, waiting for a “return message” to arrive from w. Once this return message has arrived, the node v is able to compute and send message to each of its neighbors (other than w), each being regarded, in turn, as a parent. The algorithm terminates once two messages have been passed over every edge, one in each direction. However, the Sum-Product algorithm cannot address the problems to find the state configuration that has the largest probability and calculate the corresponding marginal probability under the most likely state configuration. Moreover, as Sum-Product algorithm cannot be directly applied for factor graph model with cycles, we finally use Loopy Max-Product algorithm to address the inference tasks. Hereby, we need to modify the Sum-Product algorithm into Max-Product algorithm (Bishop, 2006) to find the state configuration $\mathrm{Y}_z^{\max}$ that maximizes the objective function $p(Y_z)$ for a specified topic z, so that:

$$Y_z^{\max} = \arg\max_{Y_z} p(Y_z) \qquad (7)$$

for which the corresponding value of the largest probability will be given by:

$$p(Y_z^{\max}) = \max_{Y_z} p(Y_z) \qquad (8)$$

Due to the cycles in our factor graph model, the proposed Loopy Max-Product algorithm first initializes the message on every link between variable node and function node in each direction as 1, and then passes messages iteratively with serial schedule using random sequences until convergence. The algorithm is summarized in Fig.2. Here update rules of the message passing for each topic z in our factor graph model can be defined as equations (9) through (12):

$$\mu_{f_{ij}^z \to y_i^z}(y_i^z) = \max_{y_j^z}[f_{ij}^z(y_i^z, y_j^z)\mu_{y_j^z \to f_{ij}^z}(y_j^z)] \qquad (9)$$

$$\mu_{f_{ij}^z \to y_j^z}\left(y_j^z\right) = \max_{y_i^z}[f_{ij}^z(y_i^z, y_j^z)\mu_{y_i^z \to f_{ij}^z}(y_i^z)] \qquad (10)$$

$$\mu_{y_i^z \to f_{ij}^z}(y_i^z) = \mu_{g_i^z \to y_i^z}(y_i^z) \prod_{f_h^z \in ne\left(y_i^z\right) \backslash f_{ij}^z, g_i^z} \mu_{f_h^z \to y_i^z}(y_i^z) \qquad (11)$$

$$\mu_{y_j^z\to f_{ij}^z}(y_j^z) = \mu_{g_j^z\to y_j^z}(y_j^z)\prod_{f_h^z\in ne(y_j^z)\backslash f_{ij}^z, g_j^z}\mu_{f_h^z\to y_j^z}(y_j^z) \qquad (12)$$

where $\mu_{f_{ij}^z\to y_i^z}(y_i^z)$ denotes the message sent from edge function node f_{ij}^z to variable node y_i^z and $\mu_{y_i^z\to f_{ij}^z}(y_i^z)$ denotes the message sent from variable node y_i^z to edge function node f_{ij}^z; $f_h^z \in ne(y_i^z)\backslash f_{ij}^z, g_i^z$ denotes the set of neighbor nodes of a given variable node y_i^z on the factor graph, excluding f_{ij}^z and g_i^z.

As every leaf node in the built factor graph is always a node function node g_i^z, its message to a variable node y_i^z is shown in equation (13). Thus, equation (11) and (12) can be further changed into equation (14) and (15):

$$\mu_{g_i^z\to y_i^z}(y_i^z) = g_i^z(y_i^z) \qquad (13)$$

$$\mu_{y_i^z\to f_{ij}^z}(y_i^z) = g_i^z(y_i^z)\prod_{f_h^z\in ne(y_i^z)\backslash f_{ij}^z, g_i^z}\mu_{f_h^z\to y_i^z}(y_i^z) \qquad (14)$$

$$\mu_{y_j^z\to f_{ij}^z}(y_j^z) = g_j^z(y_j^z)\prod_{f_h^z\in ne(y_j^z)\backslash f_{ij}^z, g_j^z}\mu_{f_h^z\to y_j^z}(y_j^z) \qquad (15)$$

So far, the maximal joint probability for the specified topic z can be obtained using equation (16) by propagating messages from the leaves to an arbitrarily chosen root node y_i^z:

$$p(Y_z)^{\max} = \max_{y_i^z}(g_i^z(y_i^z)\prod_{f_h^z\in ne(y_i^z)\backslash g_i^z}\mu_{f_h^z\to y_i^z}(y_i^z)) \qquad (16)$$

Furthermore, we can compute the marginal probability for each author by multiplying all the incoming messages as equation (17):

$$p(y_i^z) = g_i^z(y_i^z)\prod_{f_h^z\in ne(y_i^z)\backslash g_i^z}\mu_{f_h^z\to y_i^z}(y_i^z) \qquad (17)$$

Input: a citation network $G = (V, E)$, topic distribution $\{\alpha_{iz}\}_{1\le z\le t.1\le i\le N}$ and the ranking order $\{n_i\}_{1\le i\le N}$ for all authors
Output: topic-level ranking score $p(y_i^z)$ for each author $v_i \in V$
Steps:

```
1.1  Calculate each node function $g_i^z(y_i^z)$ according to Eq. (1);
1.2  Calculate each edge function $f_{ij}^z(y_i^z, y_j^z)$ according to Eq. (2);
1.3  Initialize all $\mu_{f_{ij}^z\to y_i^z}(y_i^z)$, $\mu_{f_{ij}^z\to y_j^z}(y_j^z)$, $\mu_{y_i^z\to f_{ij}^z}(y_i^z)$, $\mu_{y_j^z\to f_{ij}^z}(y_j^z)$ as 1;
1.4  Initialize all $\mu_{g_i^z\to y_i^z}(y_i^z)$ as $g_i^z(y_i^z)$;
1.5  repeat
1.6      Randomize the links in the built factor graph into a set of sequential links $L$;
1.7      for each link $l \in L$ do
1.8          if the message on $l$ is passing from an edge function node to a variable node, then
1.9              Update $\mu_{f_{ij}^z\to y_i^z}(y_i^z)$ and $\mu_{f_{ij}^z\to y_j^z}(y_j^z)$ according to Eq. (9) and Eq. (10).
1.10         end
1.11         if the message on $l$ is passing from a variable node to an edge function node, then
```

```
            Update $\mu_{y_i^z \rightarrow f_{ij}^z}(y_i^z)$ and $\mu_{y_j^z \rightarrow f_{ij}^z}(y_j^z)$ according to Eq.(14) and Eq. (15).
        end
    end
until convergence;
for each author $v_i \in V$ do
    Calculate marginal probability $p(y_i^z)$ for $v_i$ according to Eq. (17).
End
```

Fig.2 Loopy Max-Product algorithm with serial schedule using random sequences

Evaluation Methods

As this paper focuses on finding topic-based experts, it is unreasonable to directly compare our results with other classical indicators or measures for author ranking, such as H-index, citation counts, and impact factor, which are all topic-dependent. Hence we choose three topic-level baseline methods to evaluate our approach (denoted as **TWFG**), including one method that combines topic model with citation counts and two topic-based PageRank algorithms (Ding, 2011).

Topic Model & Citation Counts (TMCC)

Topic model & Citation Counts (TMCC) is used to estimate the relevance between an author and a given topic and citation counts is an important indicator for the authority of an author. For this baseline method, a subgraph of authors is first generated with a topic relevance threshold for each topic, and then the authors from the subgraph are ranked by comparing their citation counts.

Topic-based PageRank I: Simple Combination of ACT and PageRank (I_PR)

For Topic-based PageRank I (Ding, 2011), PageRank and LDA are calculated separately. Among them, the topic distributions (denoted as I) are calculated using the ACT model based on papers, where PageRank scores (denoted as PR) are calculated based on the author co-citation network. Then we simply combine ACT and PageRank as Equation (18):

$$I\_PR = \left(\frac{I-\bar{I}}{\bar{I}}\right) * \left(\frac{PR-\overline{PR}}{\overline{PR}}\right) \qquad (18)$$

where $\bar{I}$ represents the average of I and $\overline{PR}$ represents the average of PageRank.

Topic-based PageRank II: Topic-based Random Walk (PR_t)

For Topic-based PageRank II (Ding, 2011), a topic distribution is used as a weighted vector for PageRank. A topical random surfer model was proposed in which a surfer has d probability of following the links on current pages or $(1-d)\alpha$ probability of jumping to a new page, where α is the topic distribution of the new page. The topic-based PageRank II can thus be defined as Equation (19):

$$PR\_t(i) = (1-d)\frac{t(i)}{\sum_{i=1}^{N} t(i)} + d\sum_{j:j\rightarrow i}\frac{PR\_t(j)}{O(j)} \qquad (19)$$

where $t(i)$ is the conditional probability distribution of an author for a given topic and $\sum_{i=1}^{N} t(i)$ is the sum of the topic distribution of all nodes; N is the number of nodes in the network; $O(j)$ is the number of out-going links on node v_j; $PR\_t(i)$ is the topic-based PageRank on node v_i, and $PR\_t(j)$ is the topic-based PageRank on node v_j; and d is a damping factor, which is the probability that a random surfer will follow one of the links on the current page. Here, the damping factor in PR_t is set to 0.15 (to stress the equal chance of being cited), 0.50 (to indicate that scientific papers usually follow a short path of 2), or 0.85 (to stress the network topology) (Chen, Xie, Maslov & Redner, 2007).

Results and Discussion

Topic Extraction

Five topics are extracted through the ACT model, including Multimedia IR, Database and Query Processing, Medical IR, Web IR and Digital Library, and IR Theory and Model. In the meantime, topic distribution is assigned for each author. By calculating the average probability for each topic within the span of all authors, we can form a simple conclusion that Database and Query Processing is the most popular research topic out of the five topics. As shown in Table 3, we select the top 10 words to represent each extracted topic and locate the emphasis of each topic during the period from 2001 to 2008.

Table 3 Top 10 words associated with each topic

Topic No.	Topic	Word	Probability	Word	Probability
Topic 1	Multimedia IR	image	0.063250	color	0.008312
		content-based	0.017681	feedback	0.008312
		learning	0.008809	video	0.007673
		images	0.008667	semantic	0.007389
		relevance	0.008383	similarity	0.007318
Topic 2	Database and Query Processing	query	0.033203	databases	0.012764
		data	0.025732	database	0.009733
		xml	0.019248	efficient	0.009451
		processing	0.018614	web	0.009381
		queries	0.016147	querying	0.008958
Topic 3	Medical IR	database	0.010424	search	0.004138
		medical	0.007140	design	0.004138
		health	0.004982	study	0.003668
		clinical	0.004513	support	0.003575
		management	0.004325	knowledge	0.003575
Topic 4	Web IR and Digital Library	web	0.003575	system	0.005764
		search	0.015858	query	0.005764
		digital	0.008366	user	0.005607
		searching	0.006395	model	0.005212
		knowledge	0.006001	internet	0.004424
Topic 5	IR Theory and Model	document	0.014450	relevance	0.008499
		text	0.010966	fuzzy	0.008281
		query	0.009878	web	0.007991
		image	0.009587	documents	0.006829
		approach	0.008934	model	0.006539

The first topic is Multimedia IR where the main representative words are image, video, content-based, semantic, relevance, feedback, and so forth. Multimedia IR, which aims at extracting semantic information from multimedia data sources, appears as a hot topic because of its widespread sharing and exchanging of images, videos, and audio sources. Notably, content-based multimedia IR has provided new paradigms and methods for searching through the myriad variety of media throughout the world. Also, many content-based IR systems often make use of relevance feedback to refine their search results.

The second topic, Database and Query Processing, puts emphasis on the words data, XML, databases, query, web, and so forth. XML, which can encode data in a format that is both human-readable and machine-readable, has come into common use for the interchange of data over the Web. It should be pointed out that Database and Query Processing have switched from relational databases to object-oriented databases, and then further shifted to XML databases (Ding, 2011). Moreover, XML-oriented query languages for XML databases were greatly developed in order to access and manipulate XML data during the development of XML databases.

The third topic is related to Medical IR, which focuses on the words medical, health, clinical, support, knowledge, management, and so forth. As the health care industry becomes increasingly dependent on electronic information, the need to design sophisticated medical IR systems to conduct medical knowledge management also increases. In particular, as the link between health observations and health knowledge influences more choices of clinicians working toward improved health care, the clinical decision support system has been coined an “active knowledge system” to improve practitioner performance.

The forth topic focuses on Web IR and Digital Library, with digital, search, model, web, internet, and so forth as the main top words. However, Internet developments have caused most library resources to move to the Web, where related new models for publishing and searching have challenged existing methods. New digital libraries in company with Web IR have thus emerged and flourished.

The final topic, IR Theory and Model, has appeared as a popular research topic within the Information Retrieval research area. Querying and then obtaining information that meets the requirements of users from a collection of documents is the basic goal of most IR theories and models. Based on its leading words such as model, approach, web, document, query, relevance, and so forth, it would be interesting to demonstrate that researches need to adjust traditional IR models and theories to the new Web or social Web settings.

Ranking Results for Top 10 Authors

Using the proposed topical and weighted factor graph model and related inference algorithm, it is possible to provide topic-level rankings for authors. As the average probability for each author for the extracted five topics is 0.2, here we set the topic relevance threshold at $\lambda = 0.2$ with the dissimilarity threshold $\theta = 0.1$ to explicate the ranking results for the top 10 authors. Locating a small portion of experts from a large span of authors is of vital importance for sharing and propagating knowledge. As shown in Table 4, the top 10 authors based on our approach and the baseline methods for five different topics are listed. Among the results based on our approach and baseline methods, some well-known authors, such as Y. Rui and A. W. M. Smeulders (Multimedia IR), S. Abiteboul (Database and Query Processing), S. E. Robertson and A. Spink (Web IR and Digital Library), N. Fuhr (IR Theory and Model), are all ranked as among the top 10 authors. This is because these commonly ranked authors are not only highly cited, but are also relevant to the given topic. However, the top 10 authors based on these four methods are still diverse, even when several commonly ranked authors exist. It may be relevant to note that the top-ranked authors returned by PR_t(.85) are mostly the same regardless of topic. This interesting result can be explained in that PR_t(.85) focuses on network topology rather than other features encoded in the network. Another relevant observation is that I_PR is more likely to provide rankings for authors who are highly cited but not highly productive on a given topic. For example, among the top 10 authors based on I_PR, N. Fuhr (Probability = 0.09434, citation = 762), F. Crestani (Probability = 0.066667, citation = 340), and B. J. Jansen (Probability = 0.107527, citation = 651) have low probability on the Medical IR topic, but have high citation counts. Moreover, some highly relevant authors are detected by our approach but neglected by baseline methods, including C. C. Chang (Probability =0.558824, citation = 245) within the Multimedia IR topic, D. Papadias (Probability =0.58, citation = 248) within the Database and Query Processing topic, J. J. Cimino (Probability =0.367647, citation = 166) within the Medical IR topic, C. C. Yang (Probability =0.597222, citation =119) within the Web IR and Digital Library topic, and J. H. Lee (Probability =0.466667, citation = 321) within the IR Theory and Model topic.

Table 4 Top 10 authors for 5 different topics based on our approach and baseline methods

Topic	Method	Top 10 ranked authors	Method	Top 10 ranked authors
Multimedia IR	PR_t1(.85)	G. Salton, Y. Rui, J.R. Smith, S.E. Robertson, A. Spink, N.J. Belkin, T. Saracevic, E.M. Voorhees, A.W.M. Smeulders, R. Baezayates	I_PR_t1	Y. Rui, A. Spink, J.P. Eakins, T. Saracevic, J. Li, S.E. Robertson, S. Chaudhuri, J.R. Smith, R.N. Kostoff, A.W.M. Smeulders
	PR_t1(.50)	Y. Rui, J. Li, G. Salton, J.R. Smith, J.Z. Wang, N. Vasconcelos, A.W.M. Smeulders, J.P. Eakins, W.Y. Ma, S. Chaudhuri	TMCC_t1(.2)	Y. Rui, R. Baezayates, J.R. Smith, A.W.M. Smeulders, A.K. Jain, B.S. Manjunath, S. Brin, J.Z. Wang, W.Y. Ma, C Carson
	PR_t1(.15)	J. Li, N. Vasconcelos, J.Z. Wang, J.P. Eakins, S. Chaudhuri, Y. Rui, W.Y. Ma, T. Gevers, A.W.M. Smeulders, J.R. Smith	TWFG_t1(.2)	J. Li, J.Z. Wang, Y. Rui, S. Chaudhuri, A.W.M. Smeulders, J.P. Eakins, Q. Li, W.Y. Ma, N. Vasconcelos, C.C Chang
Database and Query Processing	PR_t2(.85)	G. Salton, Y. Rui, S.E. Robertson, A. Spink, N.J. Belkin, J.R. Smith, S. Abiteboul, T. Saracevic, E.M. Voorhess, D. Harman	I_PR_t2	H.V. Jagadish, A. Spink, D. Calvanese, M.J. Egenhofer, Y. Rui, G. Gottlob, T. Saracevic, S. Abiteboul, S.E. Robertson, G. Graefe
	PR_t2(.50)	G. Salton, S. Abiteboul, H.V. Jagadish, Y. Rui, J.R. Smith, S.E. Robertson, A. Gupta, G. Gottlob, N.J. Belkin, A. Spink	TMCC_t2(.2)	S. Abiteboul, T. Kohonen, Y. Yang, J. Xu, P. Buneman, A. Gupta, J. Huang, R. Fagin, S. Chaudhuri, M.J. Egenhofer
	PR_t2(.15)	H.V. Jagadish, G. Gottlob, D. Calvanese, A. Gupta, S. Abiteboul, S. Chaudhuri, M.J. Egenhofer, W.B. Frakes, L. Gravano, M. Fernandez	TWFG_t2(.2)	A. Gupta, S. Abiteboul, H.V. Jagadish , S. Santini, Y. Yang, M.J. Egenhofer, D. Papadias, G. Gottlob, D. Calvanese, W.B. Frakes
Medical IR	PR_t3(.85)	G. Salton, Y. Rui, S.E. Robertson, A. Spink, N.J. Belkin, J.R. Smith, T. Saracevic, E.M. Voorhees, D. Harman, K.S. Jones	I_PR_t3	R.N. Kostoff, A. Spink, T. Saracevic, N. Fuhr, F. Crestani, B. Hjorland, S.E. Robertson, B.J. Jansen, J.R. Smith, H.V. Jagadish

	PR_t3(.50)	G. Salton, R.N. Kostoff, Y. Rui, S.E. Robertson, N.J. Belkin, A. Spink, D.R. Swanson, J.R. Smith, T. Saracevic, S. Abiteboul	TMCC_t3(.2)	B.S. Manjuanath, C. Buckley, H. Muller, J. Xu, W. Hersh, R.N. Kostoff, C. Fellbaum, Y. Wu, R.B. Haynes, X. Lin
	PR_t3(.15)	R.N. Kostoff, H. Muller, W. Hersh, J. Li, Y. Rui, D.R. Swanson, C. Buckley, S.E. Robertson, N.J. Belkin, W.R. Hersh	TWFG_t3(.2)	R.N. Kostoff, R.B. Haynes, Y. Wu, C. Buckley, W. Hersh, J. Xu, B.S. Manjuanath, H. Muller, C.R. Shyu, J.J. Cimino
Web IR and Digital Library	PR_t4(.85)	G. Salton, A. Spink, N.J. Belkin, T. Saracevic, S.E. Roberston, Y. Rui, E.M. Voorhees, B.J. Jansen, J.R. Smith, K.S. Jones	I_PR_t4	A. Spink, T. Saracevic, B. Hjorland, S.E. Roberston, Y. Rui, B.J. Jansen, N.J. Belkin, E.M. Voorhees, R.N. Kostoff, N. Fuhr
	PR_t4(.50)	A. Spink, T. Saracevic, G. Salton, H.C. Chen, B.J. Jansen, B. Hjorland, N.J. Belkin, S.E. Robertson, P. Vakkari, E.M. Voorhees	TMCC_t4(.2)	S.E. Robertson, J.R. Smith, A. Spink, N.J. Belkin, E.M. Voorhees, T. Saracevic, B.J. Jansen, M.F. Porter, H.C. Chen, M.J. Bates
	PR_t4(.15)	A. Spink, H.C. Chen, B. Hjorland, T. Saracevic, B.J. Jansen, P. Vakkari, P. Borlund, S.E. Robertson, F. Crestani, N.J. Belkin	TWFG_t4(.2)	H.C. Chen, A. Spink, T. Saracevic, B.J. Jansen, S.E. Robertson, M. Thelwall, B. Hjorland, E.M. Voorhees, C.C. Yang, M.A. Hearst
IR Theory and Model	PR_t5(.85)	G. Salton, S.E. Robertson, A. Spink, N.J. Belkin, Y. Rui, T. Saracevic, E.M. Voorhees, N. Fuhr, J.R. Smith, K.S. Jones	I_PR_t5	N. Fuhr, T. Saracevic, C. Zhai, F. Crestani, C.J. Vanrijsbergen, A. Spink, J. Savoy, R.N. Kostoff, K.S. Jones, Y. Rui
	PR_t5(.50)	F. Crestani, G. Salton, J. Savoy, N. Fuhr, S.E. Robertson, C.J. Vanrijsbergen, A. Spink, P. Vakkarip, R. Baezayates, N.J. Belkin	TMCC_t5(.2)	S.E. Robertson, R. Baezayates, N.J. Belkin, E.M. Voorhees, N. Fuhr, B.J. Jansen, K.S. Jones, T. Kohonen, T. Joachims, C. Buckley
	PR_t5(.15)	F. Crestani, J. Savoy, N. Fuhr, P. Vakkarip, C. Zhai, C.J. Vanrijsbergen, J. Zobel, W.B Croft, H. Muller, D. Hawking	TWFG_t5(.2)	N. Fuhr, J. Savoy, F. Crestani, W.B Croft, R. Baezayates, I.J. Cox, H. Muller, D. Hawking, J. Zhang, J.H. Lee

Note: PR_t(.85), PR_t(.5), PR_t(.15) denotes topic-based random walk method with damping factor value as 0.85, 0.5, 0.15 on a given topic t, respectively; I_PR_t denotes the simple combination of the ACT model and PageRank on a given topic t; TMCC_t(.2) denotes the simple combination of the ACT model and the citation counts with a topic relevance threshold value as 0.2 on a given topic t ; TWFG denotes our topical and weighted factor graph method.

Ranking Results for Top 20 Authors and Top 50 Authors

In order to test the diversity of author rankings based on our approach and the baseline methods, we conduct a comparison of rankings for top 20 authors and top 50 authors. Here we set the topic relevance threshold as $\lambda = 0.15$ with the dissimilarity threshold $\theta = 0.1$ to calculate our ranking results as well as that of TMCC. The top 20 and top 50 authors according to TMCC are chosen as the baseline results, first calculated by topic relevance and then by citation counts. For simplicity, these top 20 and 50 authors are named as prestigious authors. Fig.3 shows the variety of rankings based on four methods for 20 prestigious authors on five different topics, and Fig.4 shows the variety of rankings based on four methods for 50 prestigious authors on five different topics. It is not surprising to see that I_PR is significantly different from other methods for all topics. The reason for this is the I_PR stress on PageRank scores, which makes the most in-link nodes overwhelming. Also, the changes in ranking results based on PR_t(.15), PR_t(.50), and PR_t(.85) are highly similar to each other because the same formulas are used for ranking, even for those with different damping factor values. In a relative sense, due to our preference for authors who are not only highly cited but also highly relevant with the given topic, our approach tends to result in very different changes compared with baseline methods.

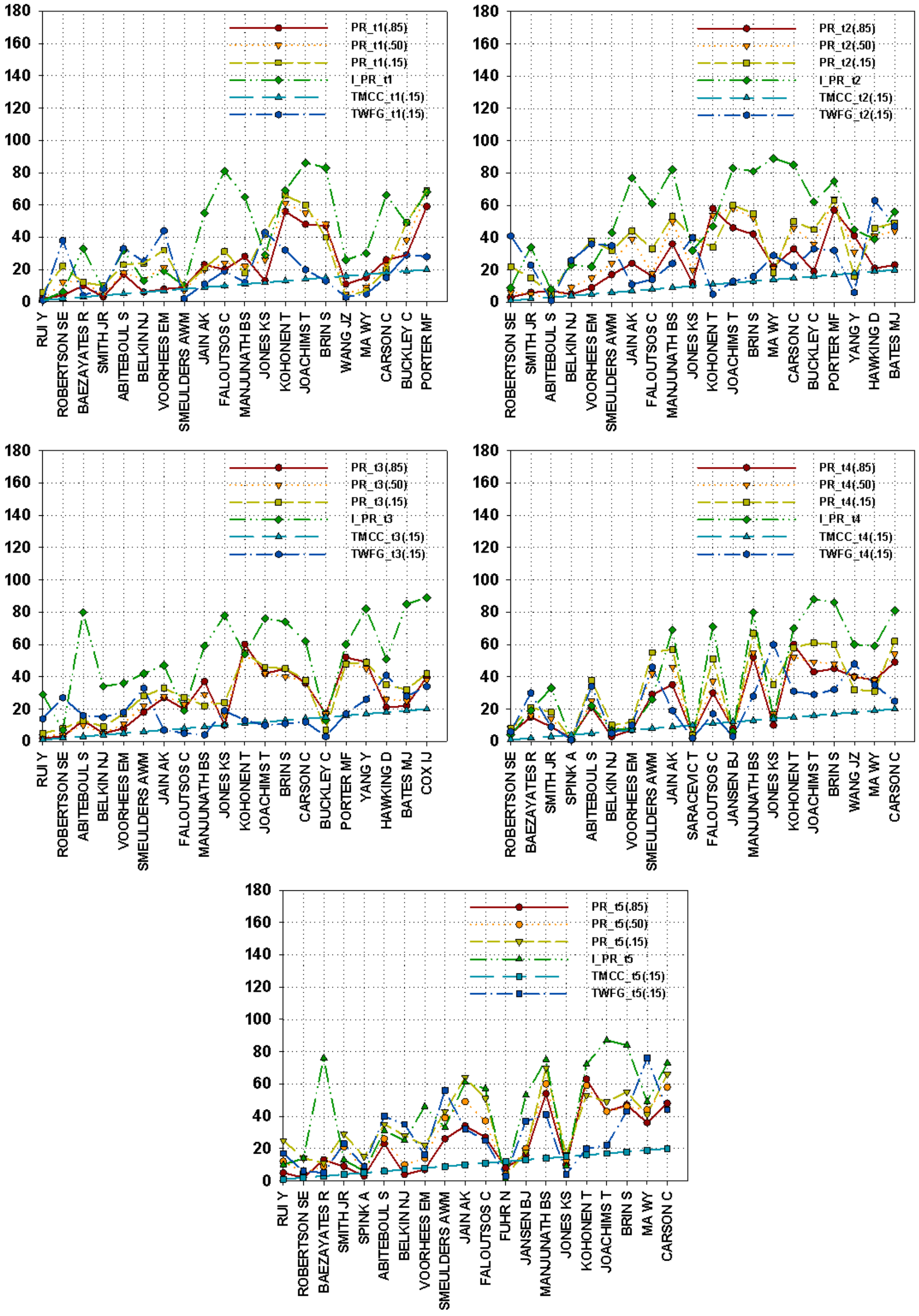

Fig.3 Rankings based on our approach and baseline methods for top 20 authors on 5 different topics (X axis represents the top 20 authors numbered based on TMCC; Y axis represents the ranks)

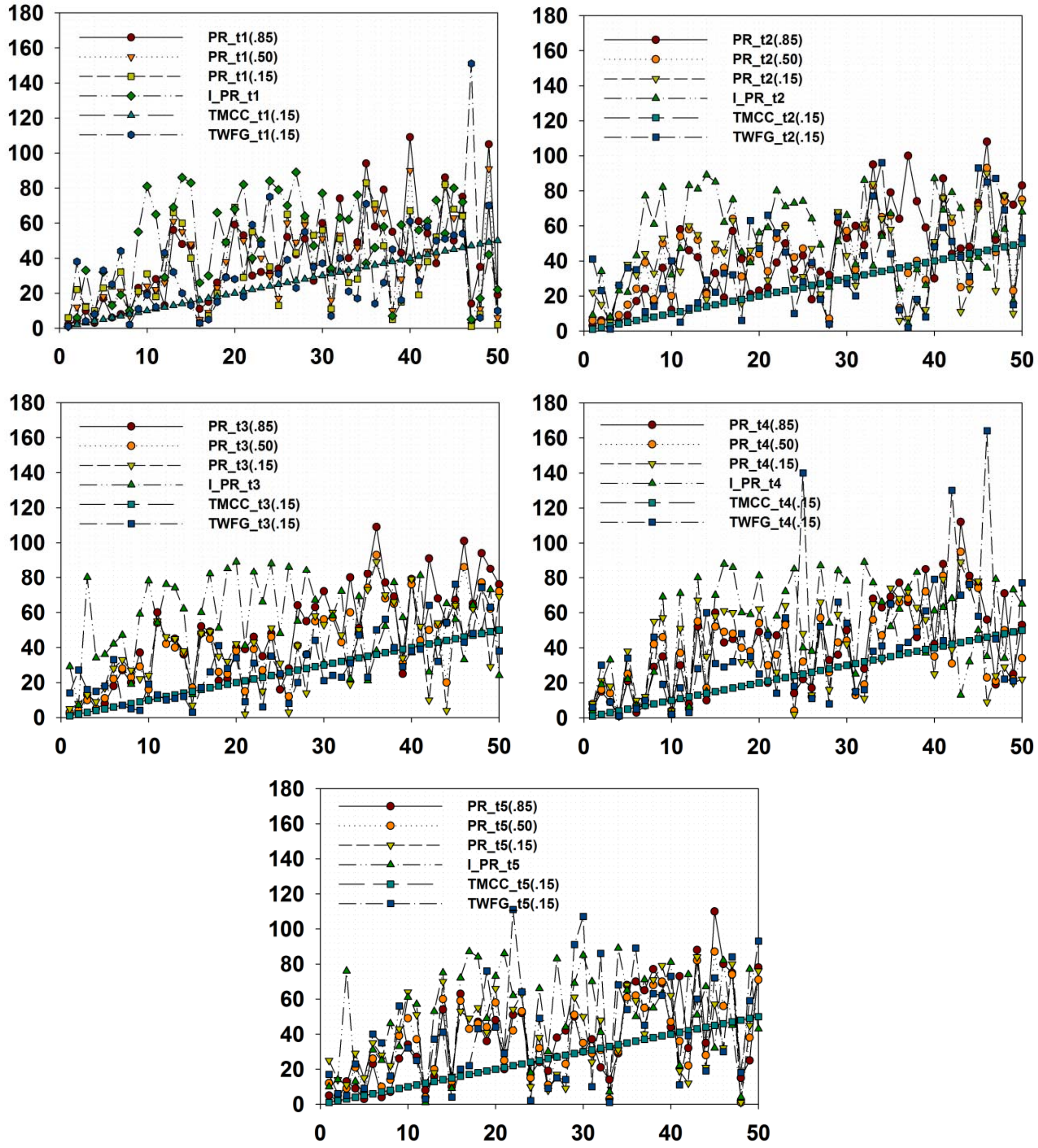

Fig.4 Rankings based on our approach and baseline methods for top 50 authors on 5 different topics (X axis represents the top 50 authors numbered based on TMCC; Y axis represents the ranks)

Evaluation

For evaluation, we use the method of pooled relevance judgments together with human judgments to generate “ground truth” from different perspectives. Assessments are first carried out in terms of topic sensitivity, and then the coverage rate of SIGIR PC members. Finally, we use the Normalized Discounted Cumulated Gain (NDCG) (Jarvelin & Kekalainen, 2002) as a metric to compare different rankings of authors based on our approach and the baseline methods.

Comparison of topic sensitivity

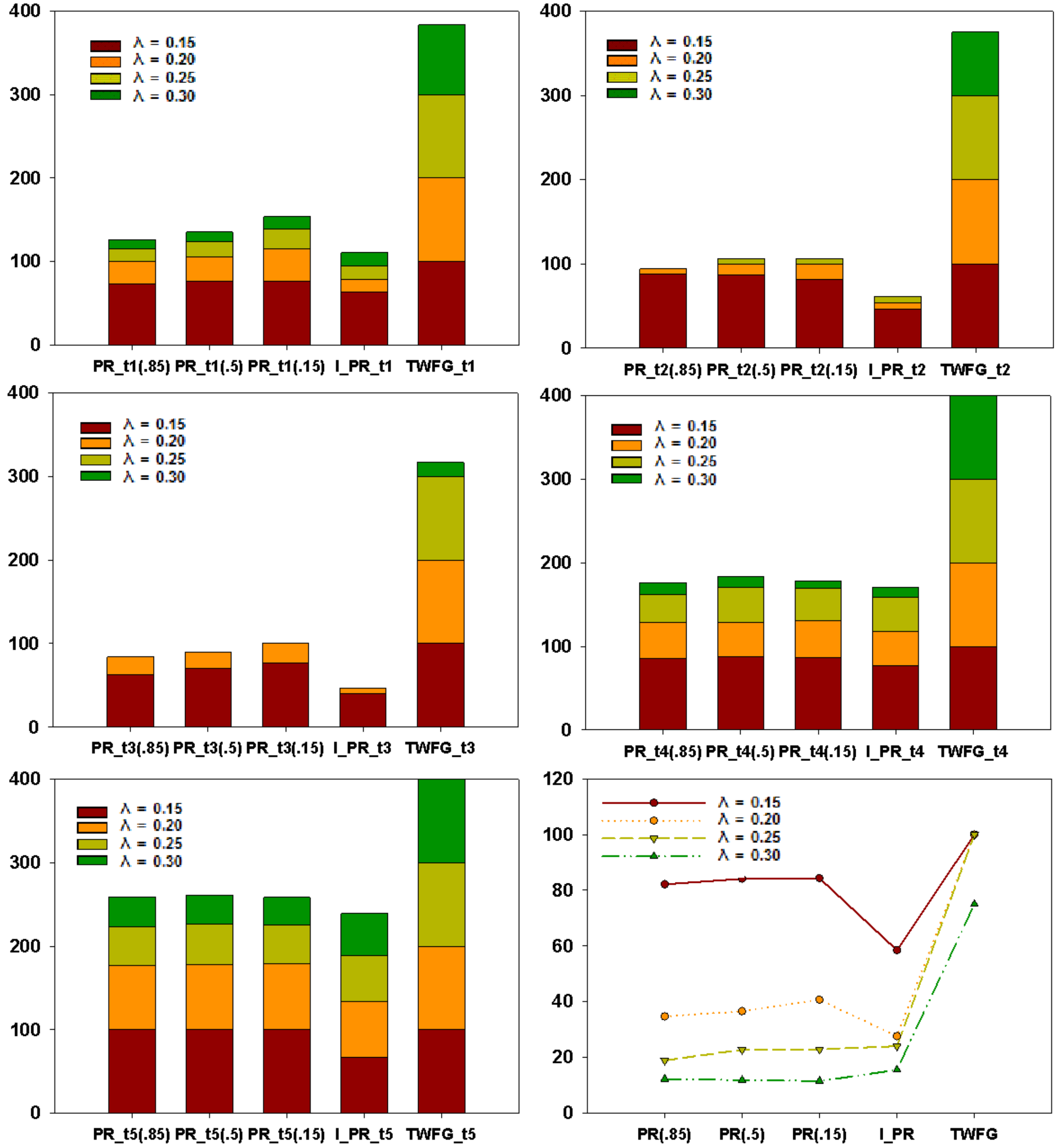

Fig.5 Comparison of topic sensitivity for each topic based on our approach, I_PR and PR_t with four different topic relevance thresholds $\lambda = 0.15, 0.2, 0.25, 0.3$

As the TMCC method is always topically sensitive, due to the fact that only authors whose topic relevance is above a defined topic relevance threshold will be chosen to rank on the top, we evaluate topic sensitivity based on our approach, I_PR, and PR_t. For most authors in the constructed citation network, their topic probabilities for five topics fall into the range from 0.15 to 0.3. Hence we select the top 100 authors together with four topic relevance thresholds, including $\lambda = 0.15$, $\lambda = 0.2$, $\lambda = 0.25$ and $\lambda = 0.3$ to demonstrate the topic sensitivity of our approach and the two other baseline methods. Based on these three methods, Fig.5 illustrates the ratio of authors whose topic probabilities exceed the value of the given topic relevance threshold for each topic, wherein the average ratio within all topics is also depicted. From the results comparison, we find that our approach, I_PR, and PR_t are comparatively topic sensitive given the low relevance threshold $\lambda = 0.15$. However, this evaluation also shows that our approach outperforms the two baseline methods, achieving an average improvement of 16.4 percent and 41.5 percent over PR_t and I_PR, even with this lower relevance threshold. Another interesting result is that I_PR and PR_t become less sensitive to a given topic when the topic relevance threshold increases, which is dramatically different from our approach. Clearly, our approach outperforms the two baseline methods in terms of topic sensitivity.

Comparison of coverage rate of the SIGIR PC members

As the ACM's Special Interest Group on Information Retrieval (SIGIR) is one of the most important international conferences for the presentation of new research results and demonstration of new systems and techniques in the field of information retrieval (IR), it is reasonable to suppose that only persons who have made significant contributions to research in information retrieval are chosen as SIGIR PC members. As suggested, a second assessment was carried out based on the professional achievement of authors selected as SIGIR PC members. We choose SIGIR PC members (i.e. Program Chairs, Program Committees, and Conference Committees) from 2001 to 2008 as the ground truth for evaluation. In order to conduct a topic-level comparison on SIGIR PC members, we tailor our evaluation data that corresponds with each topic. In other words, only SIGIR PC members whose topic probabilities are higher than a given topic relevance threshold are picked out as the ground truth on that topic. Similar to the first assessment, we conduct a comparison of the ranking results based on four topic relevance thresholds with the dissimilarity threshold $\theta = 0.1$. The result for the coverage rate of the SIGIR PC members among the top 5, 10, 20, 30, 40, and 50 with each threshold is presented in Table 5, where the best performing method is highlighted for each row. Overall, our approach achieves a better performance than the baseline methods. This evaluation shows that our approach is more effective than the baseline methods when the topic relevance threshold increases, which demonstrates our approach is more appropriate for finding professional experts within a given topic. Our approach also performs better when the coverage percentage of SIGIR PC members is higher. This result is in accord with the intuition that authors who are good at the IR Theory and Model topic more easily become experts in IR areas, and will thus have more chances to be chosen as SIGIR PC members.

Table 5 Coverage rate of the SIGIR PC members with four topic relevance thresholds within 5 topics (%)

Method	**Top@5**				**Top@10**			
	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.3$	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.3$
PR_t(.85)	24	12	8	4	34	16	12	8
PR_t(.50)	24	20	20	20	28	22	20	16
PR_t(.15)	32	32	24	24	38	34	28	24
I_PR	16	16	16	16	20	18	18	14
TMCC	24	**40**	28	**32**	28	36	**34**	28
TWFG	**36**	28	**36**	**32**	**42**	**40**	30	**32**
Method	**Top@20**				**Top@30**			
	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.3$	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.3$
PR_t(.85)	33	18	13	10	34.667	15.333	10	7.3333
PR_t(.50)	35	24	17	13	36	24	16	9.3333
PR_t(.15)	33	27	20	13	33.333	25.333	16.667	8.6667
I_PR	17	12	11	8	18	12	10.667	8
TMCC	34	**39**	35	26	**41.333**	32	32	26
TWFG	**36**	38	**36**	**33**	37.333	**36.667**	**34.667**	**32**
Method	**Top@40**				**Top@50**			
	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.3$	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.3$
PR_t(.85)	33.5	15.5	9.5	6	34.4	15.6	8.8	5.6
PR_t(.50)	35.5	20	12	7	34.8	16.8	10.4	5.6
PR_t(.15)	32	20.5	12	6.5	33.2	17.2	10	5.2
I_PR	19.5	10.5	9.5	6.5	21.6	10.4	8.8	5.6
TMCC	**39**	**33**	31.5	24.5	**37.2**	32	29.6	24.4
TWFG	33.5	32	**32**	**29**	36	**33.6**	**33.2**	**29.6**

Comparison of NDCG scores

Finally, in order to compare ranking results via the NDCG metric, a list of corresponding review/survey papers for each topic along with their citation records are recommended by the ACT model by setting a paper's topic relevance threshold at 0.2. Here we use the number of citations (named as citation score) by topic-related review papers to depict the importance of an author on the target topic. We make an assumption that if an author writes *n* papers that are separately cited by the list of review papers under a given topic, the citation score of that author for the target topic is *n*. However, for most ranked authors on each topic, the number of citations in the corresponding review papers on a given topic is less than 15. Thus if an author is cited 15 or more times by review papers on a given topic, the citation score of that author is also 15. By calculating citation scores for each author with each topic, we finally acquire the ground truth for evaluation. The comparison of NDCG scores of author rankings for each topic is demonstrated in Table 6. Overall, our approach achieves the best performance, while I_PR performs the worst with respect to NDCG scores. As the NDCG is an important measure of the average performance of a ranking algorithm, achieved by comparing the ranking results with a given "standard" ranking list, we can conclude that our approach, aimed at efficiently finding the exact experts in a given field, identifies the most authoritative experts on their topics.

Table 6 Comparison of NDCG scores of author rankings for each topic based on our approach and baseline methods

Topic	Method	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30	NDCG@40	NDCG@50
Multimedia IR	TWFG_t1(.2)	**0.5145**	**0.4465**	**0.4337**	**0.4243**	**0.4402**	**0.4551**	**0.4648**
	TMCC_t1(.2)	0.3079	0.28	0.2749	0.2982	0.33	0.3382	0.3658
	PR_t1(.15)	0.0519	0.0546	0.0608	0.0691	0.0721	0.0715	0.0502
	PR_t1(.5)	0.0633	0.0651	0.072	0.0786	0.082	0.082	0.0684
	PR_t1(.85)	0.0471	0.0498	0.0554	0.0624	0.0654	0.0642	0.0523
	I_PR_t1	0.0219	0.0246	0.0308	0.0391	0.0421	0.0415	0.0282
Database and Query Processing	TWFG_t2(.2)	**0.2142**	**0.2003**	**0.2611**	**0.2648**	**0.2594**	**0.2752**	**0.2697**
	TMCC_t2(.2)	0.068	0.074	0.0753	0.0913	0.1171	0.1223	0.1226
	PR_t2(.15)	0.05	0.0553	0.0569	0.0575	0.0569	0.0573	0.0535
	PR_t2(.5)	0.06	0.0661	0.0679	0.0683	0.0677	0.068	0.0639
	PR_t2(.85)	0.045	0.0506	0.053	0.0532	0.0526	0.053	0.0488
	I_PR_t2	0.02	0.0245	0.0263	0.0269	0.0264	0.0274	0.0236
Medical IR	TWFG_t3(.2)	**0.5002**	**0.3558**	**0.3204**	**0.3045**	**0.2878**	**0.3097**	**0.3156**
	TMCC_t3(.2)	0.1049	0.1707	0.1598	0.1619	0.1594	0.1755	0.1973
	PR_t3(.15)	0.0519	0.0535	0.0582	0.0592	0.0599	0.0594	0.0541
	PR_t3(.5)	0.0622	0.064	0.068	0.069	0.07	0.0695	0.0642
	PR_t3(.85)	0.047	0.0486	0.0531	0.0541	0.0552	0.0547	0.0493
	I_PR_t3	0.0221	0.024	0.0281	0.0294	0.0304	0.0298	0.0243
Web IR and Digital Library	TWFG_t4(.2)	**0.6712**	**0.5194**	**0.4609**	**0.4466**	**0.4241**	0.4117	0.4073
	TMCC_t4(.2)	0.5	0.5024	0.4362	0.4172	0.4033	**0.4231**	**0.4132**
	PR_t4(.15)	0.05	0.07	0.0691	0.0698	0.0702	0.0689	0.0575
	PR_t4(.5)	0.06	0.0827	0.0799	0.0802	0.0805	0.0792	0.0676
	PR_t4(.85)	0.045	0.0631	0.0626	0.064	0.0643	0.0631	0.0522
	I_PR_t4	0.02	0.043	0.0418	0.0419	0.0418	0.0405	0.0282
IR Theory and Model	TWFG_t5(.2)	**0.7617**	**0.5728**	**0.5303**	**0.5151**	**0.4768**	**0.4727**	**0.4786**
	TMCC_t5(.2)	0.3954	0.2971	0.2967	0.2881	0.3479	0.3489	0.3659
	PR_t5(.15)	0.05	0.0663	0.0671	0.0655	0.0655	0.0647	0.0566
	PR_t5(.5)	0.06	0.081	0.0806	0.0787	0.0782	0.0773	0.0677
	PR_t5(.85)	0.045	0.0617	0.0566	0.0561	0.056	0.056	0.056
	I_PR_t5	0.02	0.0392	0.0396	0.0376	0.0373	0.0364	0.0273

Conclusion and Future Work

With the emergence and rapid proliferation of social applications, the problem of expert finding has attracted increasingly more attention. Traditional methods, such as topic models, usually estimate the relevance between the candidates and a given topic but neglect the social relationships between candidates. Even link structure-based methods, such as PageRank and HITS, can be used to improve the performance of expert finding by analyzing the scholarly network information. However, to the best of our knowledge, most state-of-the-art studies tend to model the possible information on all candidates separately.

In this paper, we study the problem of topic-level expert finding in a citation network and propose a topical and weighted factor graph (TWFG) model to combine all the candidates' personal information (i.e. topic relevance and expert authority) and the scholarly network information (i.e. citation relationships) in a unified way. For topic relevance, the topic model ACT is used to extract topics and designate topic distribution for each author. And for expert authority, we use the ranking order of an author based on his/her citation counts instead of the citation count itself to represent that candidate's comprehensive knowledge. In addition, topic-level influences between neighboring candidates encoded in the citation network can offer new evidence on which to weigh their expertise on a given topic. This is accord with the fact that even with the same citation network structure, mutual influences between neighboring candidates may vary on different topics. Based on the explanations of personal and network information, the node function, edge function, and the final objective function are defined to construct our factor graph model. When conducting inference tasks on our cycle-containing factor graph model, we design the Loopy Max-Product algorithm to find the state configuration that maximizes the objective function and calculates the corresponding marginal probability for each author under the most likely state configuration.

In this paper, we choose Information Retrieval as the test field to identify experts for different topics, and compare the proposed approach with three topic-level baseline methods. In terms of topic sensitivity, it would be interesting to see whether I_PR and PR_t become less sensitive to a given topic when the topic relevance threshold increases, which is dramatically different from our approach. In addition, our approach achieves better performance than baseline methods when comparing the coverage rate of the SIGIR PC members for each topic. It should be noted that not all authors of expertise are selected as SIGIR PC members, but authors who are chosen as SIGIR PC members should be prestigious experts. Thus when the selection of SIGIR PC members becomes an important indicator for capturing professional expertise in the Information Retrieval area, it is better to use our approach. Moreover, comparison of the NDCG scores of author rankings for each topic by choosing a list of corresponding review papers as the ground truth indicates that our approach achieves the best performance. Those evaluations confirm that our factor graph-based model can definitely enhance topic-level expert-finding performance.

Future work includes identifying how to incorporate temporal information into our model in order to conduct systematical analysis of author expertise on different topics over time. Another interesting issue to pursue is combining a full-text analysis with our work to provide more precise evidence for expert finding methodology.

Acknowledgements

This work was supported by the following grants: (i) grant No. 20090094110015 from Research Fund for the Doctoral Program of Higher Education of China, (ii) grant No. BK2008354 and No. BK2010520 from the Natural Science Foundation of Jiangsu Province of China, (iii) grant No. 2008135 from the "Six Talent Peaks Program" of Jiangsu Province of China.

References

Balog, K., Azzopardi, L., & Rijke, M. D. (2006). Formal models for expert finding in enterprise corpora. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, USA, 43-50.

Balog, K., Azzopardi, L., & Rijke, M. D. (2009). A language modeling framework for expert finding. *Information Processing and Management*, *45*(1), 1-19.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer Publications, 359- 419.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1033.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45-80.

Campbell, C., Maglio, P., Cozzi, A., & Dom, B. (2003). Expertise identification using email communications. *Proceedings of 12th International Conference on Information and Knowledge Management*, New Orleans, LA, USA, 528-531.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google. *Journal of Informetrics*, 1(1), 8-15.

Ding, Y., & Cronin, B. (2010). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*(1), 80-96.

Ding, Y., Yan, E., Frazho A., & Caverlee J. (2010). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Tech*nology, *60*(11), 2229-2243.

Ding, Y. (2011). Topic-based PageRank on author co-citation networks. *Journal of the American Society for Information Science and Technology*, *62*(3), 449-466.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, *76*(1), 135-158.

Fu, Y., Xiang, R., Liu, Y., Zhang, M., & Ma, S. (2007). A CDD-based formal model for expert finding. *Proceedings of the Sixteenth Association for Computing Machinery Conference on Conference on Information and Knowledge Management*, Lisbon, Portugal, 881-884.

Gross, P.L.K., & Gross, E.M. (1927). College libraries and chemical education, *Science*, 66(1713), 385-389.

Hettich, S., & Pazzani, M. J. (2006). Mining for proposal reviewers: lessons learned at the national science foundation. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, USA, 862–871.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, USA, 50-57.

Jarvelin, K., & Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. *Association for Computing Machinery Transactions on Information Systems*, *20*(4), 422-446.

Jiao, J., Yan, J., Zhao, H., & Fan, W. (2009). ExpertRank: An expert user ranking algorithm in online communities. *Proceedings of the 2009 International Conference on New Trends in Information and Service Science*, Beijing, China, 674-679.

Jurczyk, P., & Agichtein, E. (2007). Hits on question answer portals: Exploration of link analysis for Author Ranking. *Proceedings of the 30th Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, Holland, 845-846.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, *46*(5), 604-632.

Kschischang, F. R., Frey, B. J. & Loeliger, H. (2001). Factor graphs and the sum-product algorithm. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, *47*(2), 498-519.

Kullback, S., Burnham, K. P., & Laubscher, N. F., Dallal, et al. (1987). Letter to the editor: The Kullback-Leibler distance. *The American Statistician*, *41*(4), 338-341.

Liu, X., Bollen, J., Nelson, M.L., & Sompel, H. V. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6), 1462-1480.

Matutinovic, S. F. (2007). Citation analysis for five serbian authors in web of science, scopus and google scholar. *INFOTHECA-Journal of Informatics and Librarianship*, 8 (1/2), 25-34.

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, 500–509.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the Web.* Technical Report, Stanford InfoLab, 1999-0120.

Petkova, D., & Croft, W. B. (2006). Hierarchical language models for expert finding in enterprise corpora. *Proceedings of the 18th Institute of Electrical and Electronics Engineers International Conference on Tools with Artificial Intelligence*, Washington, D.C., USA, 599-608.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Banff, Canada, 487-494.

Serdyukov, P., Henning, R., & Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. *Proceedings of the 17th Association for Computing Machinery Conference on Information and Knowledge Management*, Napa Valley, CA, USA, 1133-1142.

Sidiropoulos, A., & Manolopoulos, Y. (2006). A generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*, *79*(12), 1679-1700.

Smirnova, E., & Balog, K. (2011). A user-oriented model for expert finding. *Proceedings of the 33rd European Conference on Advances in Information Retrieval,* Dublin, Ireland, 580-592.

Smith, L. (1981). Citation Analysis. *Library Trends*, 30(1), 83-106.

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. *Proceedings of 2008 Institute of Electrical and Electronics Engineers International Conference on Data Mining,* Pisa, Italy, 1055-1060.

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. *Proceedings of the 15th Association for Computing Machinery SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 807-816.

Wu, H., Pei, Y., & Yu, J. (2009). Hidden topic analysis based formal framework for finding experts in metadata corpus. *Proceedings of the Eighth Institute of Electrical and Electronics Engineers/ACIS International Conference on Computer and Information Science*, Phoenix, AZ, USA, 369-374.

Yan, E., & Ding, Y. (2011). Discovering Author Impact: A PageRank Perspective. *Information Processing and Management*, *47*(1), 125-134.

Zhang, J., Tang J., & Li J. (2007). Expert Finding in a Social network. *Advances in Databases: Concepts, Systems and Applications,* Lecture Notes in Computer Science 4443, 1066-1069.