Identifying Topic-Specific Experts on Microblog

Yan Yu[1], Lingfei Mo[2] and Jian Wang[3]
[1] Computer Science Department, Southeast University Chengxian College
Nanjing, Jiangsu 210088 - P. R. China
[e-mail: yuyanyuyan2004@126.com]
[2] School of Instrument Science and Engineering, Southeast University
Nanjing, Jiangsu 210088 - P. R. China
[e-mail: weifanglai@sina.com]
[3] Department of Radiology, Changhai Hospital
Shanghai 200433 - P. R. China
[e-mail: wjatsh@163.com]
*Corresponding author: Yan Yu

Received December 19, 2015; revised March 16, 2016; accepted May 8, 2016; published June 30, 2016

Abstract

With the rapid growth of microblog, expert identification on microblog has been playing a crucial role in many applications. While most previous expert identification studies only assess global authoritativeness of a user, there is no way to differentiate the authoritativeness in a particular aspect of topics. In this paper, we propose a novel model, which jointly models text and following relationship in the same generative process. Furthermore, we integrate a similarity-based weight scheme into the model to address the popular bias problem, and use followee topic distribution as prior information to make user's topic distribution more precisely. Our empirical study on two large real-world datasets shows that our proposed model produces significantly higher quality results than the prior arts.

Keywords: Microblog, expert identification, topic-specific expert, LDA, similarity

A preliminary version of this paper appeared in IEEE ICC 2009, June 14-18, Dresden, Germany. This version includes a concrete analysis and supporting implementation results on MICAz sensor nodes. This research was supported by a research grant from the IT R&D program of MKE/IITA, the Korean government [2005-Y-001-04, Development of Next Generation Security Technology]. We express our thanks to Dr. Richard Berke who checked our manuscript.

http://dx.doi.org/10.3837/tiis.2016.06.010 **ISSN : 1976-7277**

1. Introduction

Microblog, such as Twitter, Sina Weibo, has become tremendously popular over the recent years. On microblog, a user follows another user, known as followee, creating an explicit following relationship. Through a formed social network which consists of users and their following relations, a user can easily broadcast a short text, known as tweet, to all of his/her followers and also automatically receive tweets from his/her followees. The rich information in microblog has become a popular resource for identifying the experts, which can be useful for many applications [1, 2], such as viral marketing, searching, expertise recommendation, information propagation, social customer relationship management etc.

A lot of studies have been done on expert identification in the context of social network. However, most of these studies, such as typical PageRank [3], only infer global authoritativeness of each user, without assessing the authoritativeness in an aspect of topics. Clearly, each user has unique topical interest and no one is an expert on every topic. Topic-specific expert analysis provides a more detailed authoritativeness portfolio for a user, which is critical for many applications [1, 2].

A few studies have been conducted to identify topic-specific experts. In general, the existing studies on topic-specific expert identification can be categorized into two camps [2]. The first camp, represented by TSPR [4] and TwitterRank [5], is PageRank-based methods. The second camp, such as Link-LDA [6] and FLDA [2], is LDA-based methods. PageRank-based methods, which extend PageRank for topical authority analysis, require the topics to be already created either manually or by a topic modeling preprocess. As the content and links are related to each other, the separation between the analysis on content and the analysis on the network structure usually leads to inferior performance, compared to LDA-based methods, which can detect topics and infer experts at the same time [2]. LDA-based methods extend Latent Dirichlet Allocation (LDA) [7], which is a popular unsupervised technique for topic discovery in large document collections. Although the LDA-based methods achieve relatively good results in topic-specific expert identification, there still exist several weaknesses that need to be addressed. First of all, some popular users on microblog are followed by many users just for their popularity. For example, President *Obama* has a massive number of followers in Twitter, but some of them are not interested in politics at all. These popular users produce a very noisy result in LDA-based model, because they repeatedly appear in almost every topic group. This phenomenon, which is called *popular bias*, makes the interpretability of topics undesirable. In the document corpus analysis, the popular users correspond to the frequent words, such as *the*, *and*, *of*, which occur in most of the documents. These frequent words do not contribute to the topic formation and also produce a very noisy result. In practice, these frequent words are manually removed before analysis according to a corpus-specific stop word list. Unfortunately, these popular users on microblog are very important to include in the analysis. Although FLDA introduces additional path-labeling process to address the popular bias problem, a followee from the popularity path is ignored and not assigned with

a topic. Secondly, LDA-based models miss the impact of user's followees in the generation of users' tweet and following relationship. Actually, users are strongly influenced by their followees on microblog. Users' followees play important roles in users' generated tweet and following relation. Therefore, the challenge of identifying topic-specific experts on microblog has yet to be studied thoroughly.

In this paper, we mainly focus on topic-specific expert identification on microblog. To address the limitations of previous approaches, we propose a novel model, which can detect topics and infer experts in the same generative process. Moreover, we address the popular bias problem by incorporating a similarity-based weight scheme into the model, and use followee topic distribution as prior information to make user's topic distribution more precisely. At last, with the inferred parameters, a search framework is introduced to produce an ordered list of topic-specific experts by their authoritativeness degree that satisfies the user's query intent.

In summary, we make the following contributions in this paper:

- We propose a new model to jointly model both tweet and following relation at the same time. Furthermore, a weight scheme is provided and the followee topic distribution is used to make user's topic distribution more precisely.
- We propose a search framework to identify topic-specific experts according to the user's query.
- Through experiments on two large real-world microblog datasets, we demonstrate that our proposed model outperforms the state-of-the-art methods in terms of accuracy.

The rest of this paper is organized as follows. In Section 2, we describe the related work. Before the details of our proposed model, we briefly review LDA model and Link-LDA model in Section 3. Section 4 introduces our model. In Section 5, we present the experimental results. Finally, we conclude the paper in Section 6.

2. Related Work

Much work has been done on expert identification in the context of social network and web structure analysis. In particular, some traditional centrality measures, which are based on the structure of social network, have been used. That is the case of Closeness and Betweenness [8, 9]. Closeness centrality is based on the length of the shortest paths from a node to everyone else. Betweenness centrality considers for each node all the shortest paths that should pass through this node to connect all other nodes in the network. PageRank [3] is a well-known algorithm used to measure both the relevance and presence of websites on the Internet, which is a variation of classic centrality measure known as eigenvector. An alternative algorithm to PageRank is HITS [10]. PageRank and HITS have been used repeatedly in the context of Twitter [11-14]. However, most of these studies only assess global authoritativeness of each user, without inferring the authoritativeness in a particular aspect of topics. In practice, topic-specific expert analysis is more effective and functional than global expert analysis for some applications.

A few studies have been conducted to identify topic-specific experts in the context of structure analysis of the web graph and social networks. In general, there are two camps on topic-specific expert identification. The first camp, represented by Topic-Sensitive PageRank (TSPR) [4] and TwitterRank [5], is PageRank-based methods. The second camp, such as Link-LDA [6] and FLDA [2], is LDA-based methods.

Given the popularity of PageRank, it is natural to extend it for topical expert analysis. TSPR is such an extension that computes per-topic PageRank scores for webpages. TSPR biases the computation of PageRank by replacing the classic PageRank's uniform teleport vector with topic-specific ones. However, it requires an existing manually categorized topic hierarchies to derive per-topic teleport vectors. TwitterRank extends TSPR to find topic-level experts on Twitter. Instead of predefined topic hierarchies, a set of topics is first produced by typical LDA on the tweets. Then TwitterRank applies a method similar to TSPR to compute the per-topic experts rank. These methods perform inferior to those approaches in the second camp that integrate text topic discovery and expert identification in the same model [2].

LDA-based methods extend the LDA, which is a popular unsupervised technique for topic discovery in large document collections. Link-LDA model is a mixed membership model to jointly model text and citations in the same generative process in the context of documents and citations. FLDA extends the Link-LDA model by capturing the content-related and content-independent reasons why a user follows another on microblog. FLDA assumes that there are two paths from which a followee can come, and introduces the path indicator to denote the path from which the followee comes. In general, although the LDA-based methods achieve relatively good results in topic-specific experts identification, there still exist several weaknesses that need to be addressed. First of all, on microblog, some popular users produce a very noisy result as they repeatedly appear in almost every topic group. Although FLDA introduced additional path-labeling process to address the popular bias problem, a followee from the popularity path is ignored and not assigned with a topic. Secondly, LDA-based models ignore the impact of user's followees in the generation of users' tweet and following relationship. In fact, users' followees play important roles in users' generated tweet and following relation.

To address the limitations of previous approaches, we propose a novel model to identify topic-specific experts on microblog.

3. LDA model and Link-LDA model

In this paper, to identify topic-specific experts on microblog, we adopt the framework of Link-LDA model, which is an extension of LDA model. For completeness, before the details of our proposed method, we first briefly review LDA model and Link-LDA model in this section.

LDA is one kind of latent topic modeling, which has become very popular as a completely unsupervised technique for topic discovery in large document collections. LDA exploits co-occurrence patterns of words in documents to unearth semantically meaningful

probabilistic clusters of words called *topics*. LDA also assigns a probabilistic membership to documents in the latent topic-space, allowing us to view and process the documents in this lower-dimensional space. In its generative process, each document is endowed with a Dirichlet-distributed vector of topic proportions, and each word of the document is assumed drawn by first drawing a topic assignment from those proportions and then drawing the word from the corresponding topic distribution. The graphical representation for LDA model is depicted in **Fig. 1**, with the notations described in **Table 1**. The generative process is summarized in **Fig. 2**.

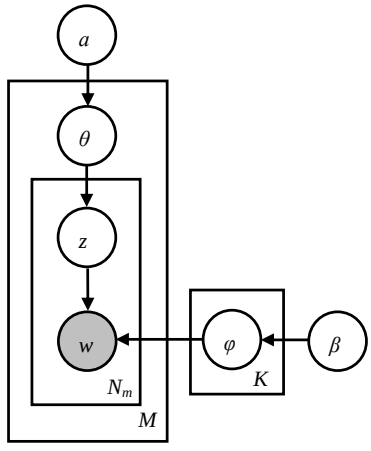

Fig. 1. Graphical representation of LDA model

Table 1. Notations used in our proposed model

Notation	Description
θ	Per-document topic distribution; Per-user topic distribution
φ	Per-topic word distribution
σ	Per-topic cited document distribution; Per-topic followee distribution
a, β, γ	Dirichlet priors on Multinominal distributions
w	Word identity
e	Cited document identity; Followee identity
z	Topic identity
M	Number of unique documents; Number of unique users
V	Number of unique words in the vocabulary
K	Number of unique topics
N_m	Number of words in document m; Number of words in the tweets of user m
L_m	Number of cited documents for document m; Number of followees for user m

For each topic k=1,..., K
  Draw $\varphi_k \sim$ Dirichlet(β)
For each document m=1,..., M
  Draw $\theta_m \sim$ Dirichlet(a)
  For the n^{th} word of the m^{th} document, where $n \in \{1,..., N_m\}$
    Draw a topic $z_{m,n} \sim$ Multinomial(θ_m), where $z_{m,n} \in \{1,..., K\}$
    Draw a word $w_{m,n} \sim$ Multinomial($\varphi_{zm,n}$), where $w_{m,n} \in \{1,..., V\}$

Fig. 2. Generative process of LDA model

Link-LDA model is an extension to LDA model. This model models text and citations in the same generative process in the context of documents and citations. It is a known fact in information retrieval that a citation between two documents not only indicates topical similarity of two documents but also authoritativeness of the cited document. In the generative process of Link-LDA, for a given document, a citation to another document is created in exactly the same way as a word is created, and they share the same per-document topic distribution. Thus, this model captures the notion that documents that share the same citations and same words, tend to be on the same topic. The document's representation in topic-space obtained from this model improves the performance of a document-classifier, compared to the representation obtained from text alone. The graphical representation for Link-LDA model is shown in **Fig. 3**. The generative process is described in **Fig. 4**.

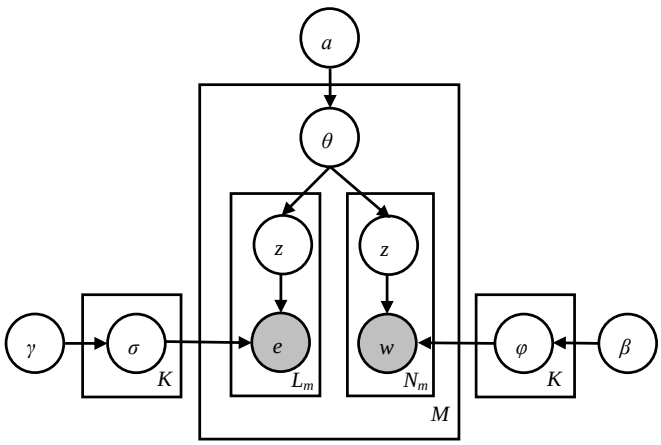

Fig. 3. Graphical representation of Link-LDA model

For each topic k=1,..., K
  Draw $\varphi_k \sim$ Dirichlet (β)
  Draw $\sigma_k \sim$ Dirichlet (γ)
For each document m=1,..., M
  Draw $\theta_m \sim$ Dirichlet (a)
  For the n^{th} word of the m^{th} document, where $n \in \{1,..., N_m\}$
    Draw a topic $z_{m,n} \sim$ Multinomial(θ_m), where $z_{m,n} \in \{1,..., K\}$
    Draw a word $w_{m,n} \sim$ Multinomial($\varphi_{zm,n}$), where $w_{m,n} \in \{1,..., V\}$
  For the l^{th} cited document of the m^{th} document, where $l \in \{1,..., L_m\}$
    Draw a topic $z_{m,l} \sim$ Multinomial(θ_m), where $z_{m,l} \in \{1,..., K\}$
    Draw a cited document $e_{m,l} \sim$ Multinomial($\sigma_{zm,l}$), where $e_{m,l} \in \{1,..., M\}$

Fig. 4. Generative process of Link-LDA model

4. Our Models

In this paper, we mainly focus on topic-specific expert identification on microblog. A topic-specific expert is defined as a user who excels in the specific topic [1]. On microblog, a user can follow another who he/she is interested and broadcast a tweet to all of his/her followers. Therefore, the tweet and following relation reflect every user's unique interest and taste. Particularly, a following relation can be interpreted as the user' vote in favor of the authoritativeness of the favorited user [1]. Therefore, we can exploit the tweet and following relation to infer the authoritativeness of a user in a specific topic.

We first introduce Base model, which models tweet and following relation in the same generative process in Section 4.1. In Section 4.2 we provide Weight model based on the Base model to address the popular bias problem. Furthermore, Followee model is described in Section 4.3, which uses followee's topic distribution as prior information to make user's topic distribution more precisely. So we get Union model by considering the similarity-based weight scheme and followee topic distribution in Section 4.4. At last, with the inferred parameters, we propose a search framework, which produces topic-specific experts that satisfy the user's query intent in Section 4.5.

4.1 Base Model

In Base model, we aggregate all the tweets for each user, and treat a user as a document and his/her following relationships to other users as citations, then Link-LDA can be applied to the microblog to identify the topic-specific experts. We assume that a user first chooses a topic from a topic distribution, and based on the chosen topic, the user chooses a word for his/her tweets. Similarly, we use the same topic distribution to choose a topic just as in the word generation. Afterwards, we choose a user to follow. Specifically, for the m^{th} user on microblog, we first pick the per-user topic distribution θ_m from a Dirichlet prior with

parameter a. Then, to generate the n^{th} word for the tweets of the user, a topic $z_{m,n}$ is first chosen from θ_m, after which the word $w_{m,n}$ is picked from the per-topic word distribution $\varphi_{zm,n}$. On the other hand, to generate the l^{th} followee for the user, we use the same topic distribution θ_m to pick a topic $z_{m,l}$ of interest, just as in the word generation part. Afterwards, we choose a followee $e_{m,l}$ who well addresses the picked topic from the per-topic followee distribution $\sigma_{zm,n}$.

By fitting the topic model to observational data, we infer the optimal values of parameters θ, φ and σ. The probabilities θ give the topic distribution for each user. The probabilities φ gives the word distribution for each topic, and the probabilities σ give the followee distribution for each topic. In particular, σ captures the likelihood of a user being followed by someone for a given topic. This value essentially quantifies the authoritativeness of a user on a given topic and is exactly the topic-specific authoritativeness score we want to compute.

Even though calculating the distributions is intractable for exact inference, various approximate inference models have been employed to estimate these distributions, including variational inference, expectation propagation, and Markov Chain Monte Carlo (MCMC) schemes. In this paper, we use Gibbs sampling [15], a special case of MCMC approximation scheme, which is widely used to approximate target distributions for LDA-like Bayesian models as it is unbiased and simple to implement. A distributed Gibbs sampling algorithm has been proposed and demonstrated excellent scalability on large clusters [2]. The posterior distributions for Gibbs sampling in Base model are given in the equations below:

$$p(z_{m,n} \mid z_{-(m,n)}, w, e; \alpha, \beta, \gamma) \propto \frac{n^{-(m,n)}_{z_{m,n},m,*} + c_{z_{m,n},m,*} + \alpha}{n^{-(m,n)}_{*,m,*} + c_{*,m,*} + T\alpha} \times \frac{n^{-(m,n)}_{z_{m,n},*,w_{m,n}} + \beta}{n^{-(m,n)}_{z_{m,n},*,*} + V\beta} \tag{1}$$

$$p(z_{m,l} \mid z_{-(m,l)}, w, e; \alpha, \beta, \gamma) \propto \frac{n_{z_{m,l},m,*} + c^{-(m,l)}_{z_{m,l},m,*} + \alpha}{n_{*,m,*} + c^{-(m,l)}_{*,m,*} + T\alpha} \times \frac{c^{-(m,l)}_{z_{m,l},*,e_{m,l}} + \gamma}{c^{-(m,l)}_{z_{m,l},*,*} + M\gamma} \tag{2}$$

where $z_{m,n}$ denotes the topic of the n^{th} word of the tweets for the m^{th} user, and $z_{m,l}$ denotes the topic of the l^{th} followee for the m^{th} user. $z_{-(m,n)}$ denotes the topic for all words and followees expect $z_{m,n}$, and $z_{-(m,l)}$ follows an analogous definition. $n_{z,m,w}$ is the number of times word w is assigned to topic z for the m^{th} user, and $c_{z,m,e}$ is the number of times followee e is assigned to topic z for the m^{th} user. * represents an aggregation on the corresponding dimension. For example, $n_{z,*,w}$ is the total number of times word w is assigned to topic z in the entire collection. $n^{-(m,n)}_{z,m,w}$ is the same meaning of $n_{z,m,w}$ only with the n^{th} word of tweets for the m^{th} user excluded. Similarly, $c^{-(m,l)}_{z,m,e}$ is defined in the same way as $c_{z,m,e}$ only without the count for the l^{th} followee for the m^{th} user.

After the sampling algorithm has run for appropriate number of iterations, the estimates for the parameters of θ, φ and σ can be obtained via the following equations:

$$\theta_{z|m} = \frac{n_{z,m,*} + c_{z,m,*} + \alpha}{n_{*,m,*} + c_{*,m,*} + K\alpha} \tag{3}$$

$$\varphi_{w|z} = \frac{n_{z,*,w} + \beta}{n_{z,*,*} + V\beta} \tag{4}$$

$$\sigma_{e|z} = \frac{c_{z,*,e} + \gamma}{c_{z,*,*} + M\gamma} \tag{5}$$

where $\theta_{z|m}$ is the probability of topic z given the m^{th} user, $\varphi_{w|z}$ denotes the probability of word w given topic z, and $\sigma_{e|z}$ represents the probability of followee e being followed by someone given topic z.

4.2 Weight Model

The two most common reasons for user's following other users are sharing a common interest and being popular on microblog [16]. Some popular users are followed by many users just for their popularity. These popular users produce a very noisy result because they repeatedly appear in almost every topic group. Recently, term weight schemes for LDA have gained intensive research interests [17, 18]. The highly frequent words are given lower weights. The results show that the term weight schemes not only lower the likelihood of highly frequent words in the topic-word distribution, but also gain a no-trivial improvement in cross-language retrieval tasks. This line of research inspires us to consider weight schemes for Base model to improve the topic interpretability in the identification of experts on microblog.

On microblog, if a user is similar with his/her follower, we may assume that this following relation between them is caused by common interest and we can assign a higher weight to the following relation. In the opposite case, we may assume that the following relation is caused merely by being popularity and we can assign a lower value to this following relation. The bigger the similarity is, the higher the influence to the topic distributions is. Therefore, we provide a Weight model by utilizing the similarity-based weight scheme to address the popular bias. In Weight model, we incorporate user's similarity as weight into Equation 1, 2, 3 and 4 by replacing the $c_{z,m,e}$ with $c^{sim}_{z,m,e}$ as follows:

$$c^{sim}_{z,m,e} = \lambda \times sim(m,e) \times c_{z,m,e} \tag{6}$$

where $c_{z,m,e}$ is the number of times followee e is assigned to topic z for the m^{th} user, $sim(m,e)$ is the similarity between the m^{th} user and his/her followee e, and λ is a scaling constant. The bigger the similarity is, the higher the influence to the topic distribution is. If we set all $sim(m,e)$ weights =1 and λ =1, this reduces immediately to Base model. The graphical representation for Weight model is shown in **Fig. 5**.

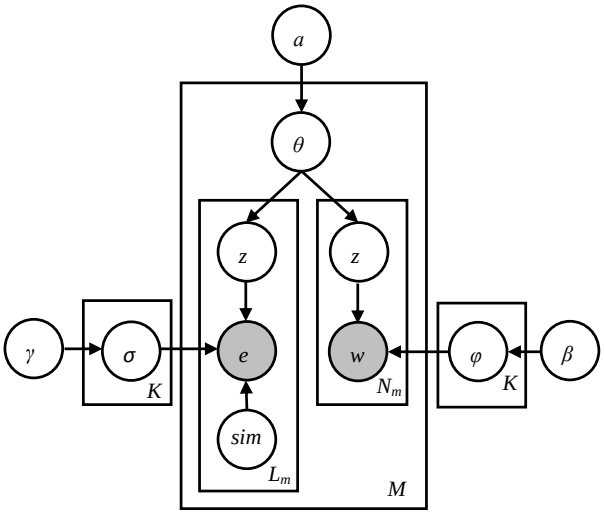

Fig. 5. Graphical representation of Weight model

User's following relation and tweet reflect user's interest and taste on microblog. Intuitively, two users can be considered similar if they share many common words in their associated tweets or follow many common users. Therefore, we utilize structure similarity and content similarity to measure users' similarity in this paper. Specifically, the content similarity of the m^{th} user and his/her followee e is defined as:

$$conSim(m,e)=\frac{\sum_{i=1}^{V} v_{m,i}v_{e,i}}{\sqrt{\sum_{i=1}^{V} v_{m,i}^2}\sqrt{\sum_{i=1}^{V} v_{e,i}^2}} \tag{7}$$

where $v_{m,i}$ represents the weight of the i^{th} word in the vocabulary for the m^{th} user. The weight of words can be calculated by using the classic TF-IDF formula. This quantity is 0 if the m^{th} user and his/her followee e have no shared words, and 1 if they have used exactly the same words, in the same relative proportions.

The structure similarity of the m^{th} user and his/her followee e is defined as:

$$struSim(m,e)=\frac{|\Gamma_+(m)\cap\Gamma_+(e)|}{|\Gamma_+(m)\cup\Gamma_+(e)|} \tag{8}$$

where $\Gamma_+(m)$ is the set of followees of the m^{th} user, $|\cdot|$ denotes the size of the set.

We make a linear combination of the content similarity and structure similarity. As a result, the similarity between the m^{th} user and his/her followee e is:

$$sim(m,e)=conSim(m,e)+struSim(m,e) \tag{9}$$

4.3 Followee Model

Users are strongly influenced by their followees on microblog. Base model however ignores the impact of user's followees in the generation of users' tweets and following relations. In order to reflect the intuition that followees have impact on the content constitution of the users, we provide Followee model by integrating followees into Base model as prior information to make user's topic distribution more precisely. In the generation of tweets and followees, topic of a user is split into two parts: the idea of the user and the knowledge from his/her followees. We combine user and his/her followees' topic distributions together to generate the topic. So the generating the topic layer is no longer controlled only by the user topic distribution only. Instead, both the user and his/her followees play the role of generating the topic. However, among all of a user's followees, some followees may have similar taste with this user, while some other followees may have different tastes. We use similarity which allows us to treat user's followees differently. If the m^{th} user and his/her followee e are very similar, then followee e should contribute more. On the other hand, if these two users are dissimilar, then e should contribute less. Thus, the combined topic distribution is:

$$\theta_m^c = \theta_m + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times \theta_e}{\sum_{e \in \Gamma_+(m)} sim(m,e)} \tag{10}$$

Fig. 6 shows the graphical representation of Followee model, and **Fig. 7** describes its generative process.

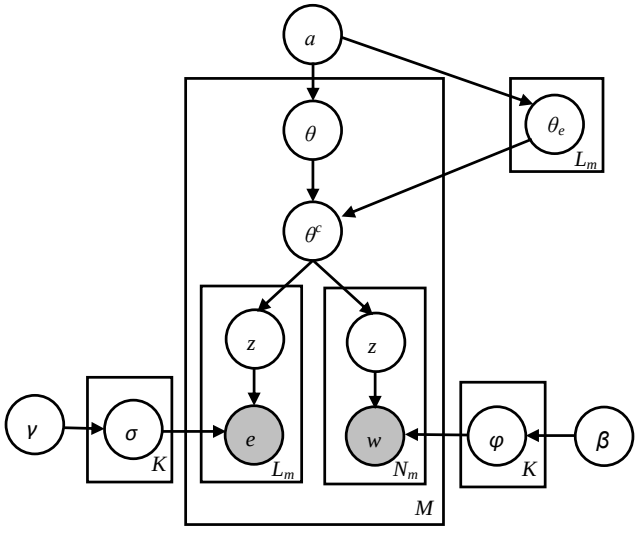

Fig. 6. Graphical representation of Followee model

For each topic k=1,...,K

Draw φ_k ~ Dirichlet(β)

Draw σ_k ~ Dirichlet(γ)

For each user m=1,...,M

Draw θ_m ~ Dirichlet(a)

For each followee e of the m^{th} user, where $e \in \{1,..., L_m\}$

Draw θ_e ~ Dirichlet(a)

Combine θ_m and θ_e to generate a combined topic distribution θ_m^c

For the n^{th} word of the m^{th} user, where $n \in \{1,..., N_m\}$

Draw a topic $z_{m,n}$ ~ Multinomial(θ_m^c), where $z_{m,n} \in \{1,..., K\}$

Draw a word $w_{m,n}$ ~ Multinomial($\varphi_{zm,n}$), where $w_{m,n} \in \{1,..., V\}$

For the l^{th} followee of the m^{th} user, where $l \in \{1,..., L_m\}$

Draw a topic $z_{m,l}$ ~ Multinomial(θ_m^c), where $z_{m,l} \in \{1,..., K\}$

Draw a followee $e_{m,l}$~ Multinomial($\sigma_{zm,l}$), where $e_{m,l} \in \{1,..., M\}$

Fig. 7. Generative process of Followee model

The posterior distributions for Gibbs sampling in Followee model are:

$$p(z_{m,n} \mid z_{-(m,n)}, w, e; \alpha, \beta, \gamma)$$

$$\propto \frac{n_{z_{m,n},m,*}^{-(m,n)} + c_{z_{m,n},m,*} + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times (n_{z_{m,n},e,*} + c_{z_{m,n},e,*})}{\sum_{e \in \Gamma_+(m)} sim(m,e)} + \alpha}{n_{*,m,*}^{-(m,n)} + c_{*,m,*} + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times (n_{*,e,*} + c_{*,e,*})}{\sum_{e \in \Gamma_+(m)} sim(m,e)} + K\alpha} \tag{11}$$

$$\times \frac{n_{z_{m,n},*,w_{m,n}}^{-(m,n)} + \beta}{n_{z_{m,n},*,*}^{-(m,n)} + V\beta}$$

$$
\begin{aligned}
& p(z_{m,l} \mid z_{(-m,l)}, w, e; \alpha, \beta, \gamma) \\
& \propto \frac{n_{z_{m,l},m,*} + c^{-(m,l)}_{z_{m,l},m,*} + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times (n_{z_{m,l},e,*} + c_{z_{m,l},e,*})}{\sum_{e \in \Gamma_+(m)} sim(m,e)} + \alpha}{n_{*,m,*} + c^{-(m,l)}_{*,m,*} + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times (n_{*,e,*} + c_{*,e,*})}{\sum_{e \in \Gamma_+(m)} sim(m,e)} + K\alpha} \\
& \times \frac{c^{-(m,l)}_{z_{m,l},*,e_{m,l}} + \gamma}{c^{-(m,l)}_{z_{m,l},*,*} + M\gamma}
\end{aligned}
\tag{12}
$$

Then parameters θ, φ and σ are estimated as follow:

$$
\theta_{z|m} = \frac{n_{z,m,*} + c_{z,m,*} + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times (n_{z,e,*} + c_{z,e,*})}{\sum_{e \in \Gamma_+(m)} sim(m,e)} + \alpha}{n_{*,m,*} + c_{*,m,*} + \frac{\sum_{e \in \Gamma_+(m)} sim(m,e) \times (n_{*,e,*} + c_{*,e,*})}{\sum_{e \in \Gamma_+(m)} sim(m,e)} + K\alpha}
\tag{13}
$$

$$
\varphi_{w|z} = \frac{n_{z,*,w} + \beta}{n_{z,*,*} + V\beta}
\tag{14}
$$

$$
\sigma_{e|z} = \frac{c_{z,*,e} + \gamma}{c_{z,*,*} + M\gamma}
\tag{15}
$$

4.4 Union Model

In this section, we design Union model to model similarity-based weight scheme and followee-based prior information into the Base model simultaneously to improve the accuracy of expert identification on microblog. The posterior distributions and parameters in Union model are obtained by replacing $c_{z,m,e}$ in Equation 11, 12, 13, and 15 with $c^{sim}_{z,m,e}$.

4.5 Querying Topic-Specific Authorities

Finally, we propose a search framework for topic-specific expert identification on microblog. The framework allows a user to express his/her interests by typing a set of keywords. Then the framework returns an ordered list of experts by their authoritativeness score that satisfy the user's intent.

More specifically, by fitting the topic model to observational data, the optimal values of parameters $\theta_{z|m}$ and $\sigma_{e|z}$ are inferred as part of the result. $\theta_{z|m}$ is the probability of topic z given the m^{th} user, and $\sigma_{e|z}$ represents the probability of followee e being followed by someone given topic z. If we treat a query q as a new user, we can learn $\theta_{z|q}$ from our model,

the probability of topic z given the query q, which represents the weight of interested topics of the query keyword. On the other hand, $\sigma_{e|z}$ can be quantified the authoritativeness of followee e on a given topic z. At the same time, the LDA-based model is too coarse to be used as the only representation [19]. So we combine the content similarity to refine the query. As the result, the final authoritativeness score $auth(q,u)$ for a user u given a query q is computed as

$$auth(q,u) = (1 + conSim(q,u)) \times \sum_{k=1}^{K} (\theta_{k|q} \times \sigma_{u|k}) \tag{16}$$

Finally, the experts are returned in decreasing order of their authoritativeness scores $auth(q,u)$.

5. Experiments

In this section, we evaluate the effectiveness of our proposed model on two real-world microblog datasets collected from Tencent Weibo and Sina Weibo. We quantitatively analyze our proposed model on Tencent Weibo dataset in Section 5.1. We then give examples of topic-specific experts on Sina Weibo dataset in Section 5.2.

5.1 Effectiveness on Tencent Weibo Dataset

In this section, we systematically evaluate the effectiveness of our proposed model on a dataset from Tencent Weibo. Tencent Weibo is a Chinese microblog website, launched by Tencent in 2010. It has become one of leading microblog platforms in China. The dataset we use for evaluation in this paper is the dataset used in the KDD Cup 2012 Track1. Track 1 in KDD Cup 2012 provides rich information across multiple domains such as user profiles, following relationship, and keyword. In particular, the dataset contains the set of provided VIP users, which enables us to systematically evaluate the accuracy of various methods. These VIP users are manually labeled by Tencent Weibo administrators, According to Tencent Weibo, the VIP users are typically famous people and organizations. In other words, they are "experts" in their corresponding topics. As a result, the VIP users can be used as the "ground truth" for our empirical evaluation. The basic statistics of the Tencent Weibo dataset is given in **Table 2**. We held out 10% of the data for test purposes and trained the models on the remaining 90%.

Table 2. Statistics of experimental datasets

Dataset	#users	# dist. words	#total words	#following relationship
Tencent Weibo	2.33M	714K	492M	51M
Sina Weibo	1.78M	36M	105M	311M

We first evaluate the quality of the discovered topics based on two evaluation metrics: avg_catSim and avg_KL. In the topic model, σ captures the likelihood of a user being followed by someone for a given topic. This value can be viewed as the topic-specific authoritativeness score. Therefore, we choose the top 50 users for each topic as

topic-specific experts in decreasing order of their values σ, and measure the category similarity of the top 50 users. In this dataset, VIP users are organized in hierarchical categories, where categories in different levels are separated by a dot, and categories are anonymized as integers, such as 1.2.4.5. Formally, we suppose that vector $G_u=(g_{u,1},g_{u,2},...,g_{u,t})^T$ represents the hierarchical category for VIP user u, where $g_{u,i}$ is the category in the i^{th} level for user u, and t is the number of category level. The category similarity of user u and user v in topic z is defined as:

$$catSim(u,v,z) = log_2((\frac{\sum_{i=1}^{t} g_{u,i} \oplus g_{v,i}}{t} + \varepsilon) \times \sigma_{u|z} \times \sigma_{v|z}) \tag{17}$$

where ε is a small number to avoid logarithm of zero, and $\oplus$ is defined as:

$$g_{u,i} \oplus g_{v,i} = \begin{cases} 1 & if\ g_{u,i} = g_{v,i} \\ 0 & otherwise. \end{cases} \tag{18}$$

avg_catSim is the average value of catSim of all topics. By definition, a higher avg_catSim score indicates a better algorithm.

Kullback-Leibler divergence (D_{KL}) is also known as the relative entropy, which is generally used to reflect the difference between two probability distributions. The average Kullback-Leibler divergence (avg_KL) between topics can also expressed as the performance of the algorithm. A higher avg_KL score indicates a better algorithm. Formally, avg_KL is defined as:

$$avg\_KL = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} D_{KL}(\sigma_i \,||\, \sigma_j)}{K^2} \tag{19}$$

$$D_{KL}(\sigma_i \,||\, \sigma_j) = \sum_{u=1}^{T} \sigma_{u|i} log \frac{\sigma_{u|i}}{\sigma_{u|j}} \tag{20}$$

where T is the total number of VIP users.

We compare Base, Weight, Followee and Union model, which are described in Section 4.1, 4.2, 4.3 and 4.4 respectively. For every algorithm, we set a=0.1, β=0.1 and γ=0.1. The number of topics K is varied from 20 to 100. We run Gibbs sampling for 500 iterations. We run multiple runs to find the optimal parameter value λ=30.

Fig. 8 and **Fig. 9** present the avg_catSim and avg_KL of the discovered topics as a function of the numbers of topics for four models on Tencent Weibo dataset. As shown in both figures, Base model is usually the worst among all methods. The Weight model and Followee model are consistently better than Base model by a significant margin. The Weight model extends Base model by incorporating users' similarity as weight into the conditional probabilities to address the popular bias. Followee model views the followees as the prior information which users have had to refine the users' topic distribution. The results are consistent with our intuition. Union model produces better results than the others, thanks to simultaneously incorporating the similarity and prior information of followees to further refine the model.

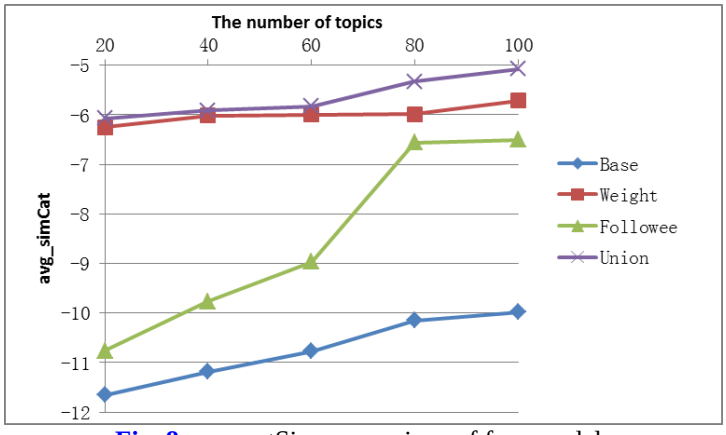

Fig. 8. avg_catSim comparison of four models

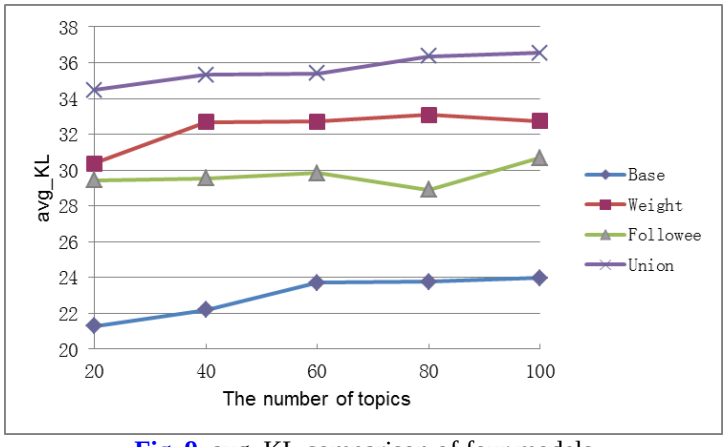

Fig. 9. avg_KL comparison of four models

We then measure the accuracy of querying topic-specific experts of our proposed model. For each category, we use one VIP user (i.e. all the keywords of this user) as the query, and employ the standard Mean Reciprocal Rank (MRR) metric to analyze the results across all categories. Let Q denote a set of queries. For each query $q \in Q$, each algorithm returns an ordered list of users by their authoritativeness. The Reciprocal Rank of a ranked list is the multiplicative inverse of the rank of the first hit in the list. The MRR score of an algorithm is the average reciprocal rank obtained by the ranked lists given by the algorithm with respect to the query set Q. Formally,

$$MRR = \frac{1}{|Q|}\sum_{q \in Q}\frac{1}{rank_q} \tag{21}$$

where $rank_q$ is the rank of the first real expert in the ranked list for query q. By definition, a higher MRR score denotes a better algorithm.

In the following, we compare Union model with TwitterRank and FLDA. The

TwitterRank algorithm was originally proposed to find topic-level authorities on Twitter. It extends typical Topic-Sensitive PageRank to compute per-topic influence scores. The transition probability between two nodes in TwitterRank is defined based on the topical similarity between the corresponding users. FLDA can detect topics and infer experts at the same time. Furthermore, it differentiates the different reasons why a user follows another. In our experiments, we set the number of topics to 100, and run Gibbs sampling for 500 iterations. The priors used are 0.1 for a, β and γ. For Union model, λ is 30.

Fig. 10 shows the MRR score of each algorithm on the Tencent Weibo dataset. It is observed that TwitterRank is inferior to the other two algorithms because of separation between the topic analysis and the expert detection. FLDA is inferior to Union model. Although FLDA introduces additional path-labeling process to address the popular bias, a followee from the popularity path is ignored and not assigned with a topic. Union model outperforms all the other algorithms for detecting topics and inferring experts in the same time, and incorporating the similarity and prior information of followee.

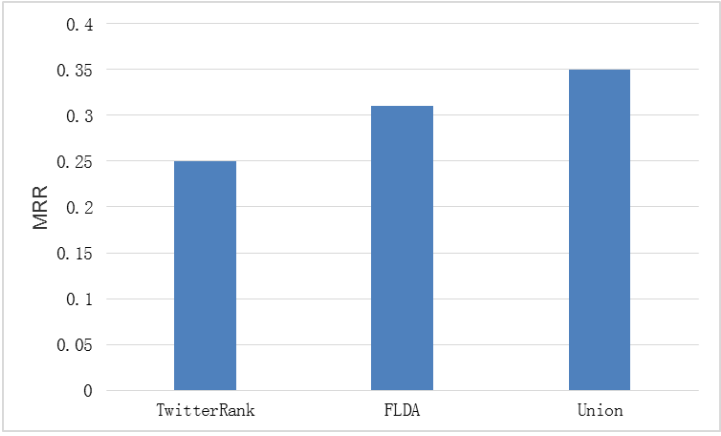

Fig. 10. MRR comparision of three models

5.2 Effectiveness on Sina Weibo Dataset

In this section, we give examples of search on Sina Weibo dataset for a list of experts. Sina Weibo is the largest microblog service in China. The dataset we use in this section is published by Jie Tang's group (https://aminer.org/Influencelocality), which is crawled from Sina Weibo between October 2009 and January 2010. The dataset contains users' descriptions, following relationships among them, and their tweets. The basic statistics of the dataset is given in **Table 2**. We use 80% of the data as a training set and the remaining 20% of the data as a test set. We compare Union model with TwitterRank and FLDA. We set a=0.1, β=0.1, γ=0.1, λ=30 and K=100. We run Gibbs sampling for 500 iterations.

Table 3 shows the results of top five experts identified by TwitterRank, FLDA and Union model in two specific topics respectively. Together with Sina Weibo User ID, we show rank, follower count, and the description of each user. By going over the users' descriptions, we highlight the irrelevant experts for clarity. As shown from the table, we

can see that Union model produces better results than the competing methods in two specific topics. For example, for the first query topic: *delicious food*, the second user and the third user in the ranked list identified by TwitterRank are not relevant to *delicious food*. FLDA misidentified an Internet analyst as expert in *delicious food*. By contrast, Union model successfully identified popular users relevant to this specific topic. For the second query topic: *photography*, Car enthusiast found by TwitterRank and Chinese Basketball News produced by FLDA are much less relevant to *photography*, whereas the top five users produced by Union model are all relevant to *photography*. From the results of the table, we note that TwitterRank produces more irrelevant experts with high numbers of followers, which clearly shows that separating the topic discovery and expert identification performs inferior to the approaches that integrate them in the same process. FLDA model reduces the chance of popular users' appearing in the irrelevant topic groups by labeling non-topic-driven following relations with popularity path. However, FLDA ignores topics of some popular users. As a result, some popular relevant users are removed from the specific topics. At the same time, FLDA still suffers from presence of a few irrelevant users in the specific topics. Compared with TwitterRank and FLDA, Union model produced better results by incorporating the similarity and prior information of followees in Base model to produce the user's topic distribution more precisely.

Table 3. Examples of top five experts identified by three models

Query Topic: *delicious food*				
Model	**UserID**	**Rank**	**#follower**	**Description**
TwitterRank	1669763744	1	4,718,609	Gourmet, writer, movie producer
	1931358544	2	4,979,789	Film website
	1248351970	3	2,483,891	Beauty consultant
	1024763102	4	1,002,048	Food program
	1250369822	5	531,100	Expert of food website
FLDA	1571343005	1	105,322	Program host of cooking and traveling
	1442923510	2	109,958	Food magazine producer
	1752640191	3	58,035	Chief chef
	1770277210	4	107,397	Recipe website
	1935871593	5	114,139	Internet analyst
Union	1669763744	1	4,718,609	Gourmet, writer, movie producer
	2188014311	2	1,279,876	Food website
	1684322821	3	611,245	Cooking program host
	1622749134	4	255,467	Food writer, media person
	1384615417	5	628,271	Home cuisine magazine
Query Topic: *photography*				
Model	**User ID**	**Rank**	**#follower**	**Description**
TwitterRank	1883617103	1	1,014,498	Photographer, writer
	2119408713	2	816,538	Traveler, photographer
	1845322103	3	417,482	Car enthusiast
	1807646700	4	345,281	Photography Forum
	1220824437	5	735,104	Food blogger, photographer

FLDA	1276539303	1	122,022	Musician, photographer, director.
	1684866844	2	26,650	Photographic studio
	1050979560	3	141,863	Traveling and photographing blogger
	1737961042	4	501,565	Chinese Basketball News
	1669193880	5	15,302	Photographer
Union	1883617103	1	1,014,498	Photographer, writer
	1807646700	2	345,281	Photography Forum
	1232204102	3	335,794	Photographer, traveler and geographer
	1784467952	4	275,370	POCO official photography community
	1218149847	5	424,595	Fashion photographer

6. Conclusion

This paper addresses the problem of topic-specific expert identification on microblog. To model topic-specific authoritativeness, we introduce a novel method, which jointly models text and following relation in the same generative process. Furthermore, we integrate a similarity-based weight scheme into the model and use followee topic distribution as prior information to make user's topic distribution more precisely. Our empirical study on two real-world datasets shows that our proposed model produces higher quality results than the prior arts.

Acknowledgements

This research is supported by the Research on University’s Natural Science Projects of Jiangsu Province (No. 14KJB520004).

References

[1] Bin Bin, Ben Kao, Chang Wan and Junghoo Cho, "Who are experts specializing in landscape photograph?: analyzing topic-specific authority on content sharing services," in *Proc. of 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1506-1515, August 24-27, 2014. Article (CrossRef Link)

[2] Bin Bi, Yuanyuan Tian, Yannis Sismanis, Andrey Balmin and Junghoo Cho, “Scalable topic-specific influence analysis on microblogs,” in *Proc. of the 7th ACM International Conference on Web Search and Data Mining*, pp. 513-522, February 24-28, 2014. Article (CrossRef Link)

[3] Sergy Brin and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine,” in *Proc. of the Seventh International World Wide Web Conference*, pp. 107-117, April 14-18, 1998. Article (CrossRef Link)

[4] Taher H. Haveliwala, “Topic-sensitive PageRank,” in *Proc. of the 11th International Conference on World Wide Web*, pp. 517-526, May 7-11, 2002. Article (CrossRef Link)

[5] Jianshu Weng, Ee-Peng Lim, Jing Jiang and Qi He, “TwitterRank: finding topic-sensitive influential twitterers,” in *Proc. of the third ACM International Conference on Web Search and Data Mining*, pp. 261-270, February 4-6, 2010. Article (CrossRef Link)

[6] Elena Erosheva, Stephen Fienberg and John Lafferty, "Mixed-membership models of scientific publications," in *Proc. of the National Academy of Sciences of the United States of America*, pp. 5220-5227, April 6, 2004. Article (CrossRef Link)
[7] Thomas Hofmann, "Probabilistic latent semantic indexing," in *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57, August 15-19, 1999. Article (CrossRef Link)
[8] Behnam Hajian and Tony White, "Modelling influence in a social network: metrics and evaluation," in *Proc. of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 497-500, October 9-11, 2011. Article (CrossRef Link)
[9] Xin Jin and Yaohua Wang, "Research on social network structure and public opinions dissemination of micro-blog based on complex network analysis," *Journal of Networks*, vol. 8, no. 7, pp. 1543-1550, July, 2013. Article (CrossRef Link)
[10] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604-632, September, 1999. Article (CrossRef Link)
[11] Haewoon Kwak, Changhyun Lee, Hosung Park and Sue Moon, "What is Twitter, a social network or a news media?," in *Proc. of the 19th International Conference on World Wide Web*, pp.591-600, April 26-30, 2010. Article (CrossRef Link)
[12] Chun Chen, Feng Li, Beng Chin Ooi and Sai Wu, "TI: an efficient indexing mechanism for real-time search on tweets," in *Proc. of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 649-660, June 12-16, 2011. Article (CrossRef Link)
[13] Shoubin Kong and Ling Feng, "A tweet-centric approach for topic-specific author ranking in micro-blog," in *Proc. of The 7th International Conference on Advanced Data Mining and Applications*, pp. 138-151, December 17-19, 2011. Article (CrossRef Link)
[14] Lamjed Ben Jabeur, Lynda Tamine and Mohand Boughanem, "Active microbloggers: identifying influencers, leaders and discussers in microblogging networks," in *Proc. of The 19th International Symposium on String Processing and Information Retrieval*, pp. 111-117, October 21-25, 2012. Article (CrossRef Link)
[15] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," in *Proc. of the National Academy of Sciences of the United States of America*, pp. 5228-5235, April 6, 2004. Article (CrossRef Link)
[16] Youngchul Cha and Junghoo Cho, "Social-network analysis using topic models," in *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 565-574, August 12-16, 2012. Article (CrossRef Link)
[17] Xiaona Wu, Jia Zeng, Jianfeng Yan and Xiaosheng Liu, "Finding better topics: features, priors and constraints," in *Proc. of The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 296-310, May 13-16, 2014. Article (CrossRef Link)
[18] Youngchula Cha, Bin Bi, Chu-Cheng Hsieh and Junghoo Cho, "Incorporating popularity in topic models for social networks analysis," in *Proc. of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 223-232, July 28-August 1, 2013. Article (CrossRef Link)
[19] Uan Cao, Jintao Li, Yongdong Zhang and Sheng Tang, "LDA-based retrieval framework for semantic news video retrieval," in *Proc. of First IEEE International Conference on Semantic Computing*, pp. 155-160, September 17-19, 2007. Article (CrossRef Link)

Yan Yu is an associate professor at Southeast University Chengxian College, Nanjing, Jiangsu, P. R. China. She has a PhD in Computer Science from Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, P. R. China. Her current research interests lie in the study of machine learning, graphical models, information extraction and social network.

Lingfei Mo is an associate professor at Southeast University, Nanjing, Jiangsu, P. R. China. He received his PhD degree from Zhejiang University, Hangzhou, Zhejiang, P. R. China. His current research interests include machine learning, logic programming, reasoning under uncertainty, and computer aided diagnosis.

Wang Jian is a medical doctor, radiologist and associate professor of department of radiology of Changhai Hospital, Second Military Medical University, Shanghai, P. R. China. He received his bachelor of medicine degree from Second Military Medical University, Shanghai, P.R.China in 1996, and a PhD degree from the same university in 2004. His current research interests include medical picture archiving and communication, large volume of medical picture processing and computer aided diagnosis.