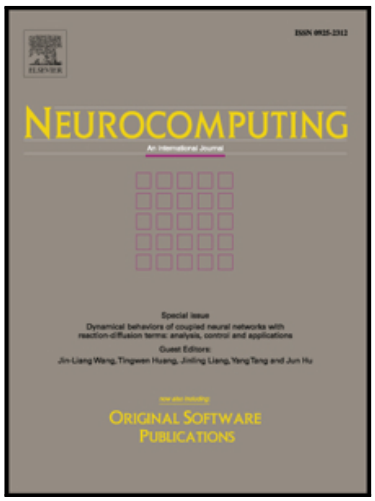
Accepted Manuscript



Combining Paper Cooperative Network and Topic Model for Expert Topic Analysis and Extraction

Shengxiang Gao , Xian Li , Zhengtao Yu , Yu Qin , Yang Zhang

PII: S0925-2312(17)30159-5
DOI: 10.1016/j.neucom.2016.12.074
Reference: NEUCOM 17971

To appear in: *Neurocomputing*

Received date: 15 July 2016
Revised date: 27 November 2016
Accepted date: 3 December 2016

Please cite this article as: Shengxiang Gao , Xian Li , Zhengtao Yu , Yu Qin , Yang Zhang , Combining Paper Cooperative Network and Topic Model for Expert Topic Analysis and Extraction, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.12.074

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Combining Paper Cooperative Network and Topic Model for Expert Topic Analysis and Extraction

Shengxiang Gao, Xian Li, Zhengtao Yu*, Yu Qin, Yang Zhang

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500)

Corresponding Author: Zhengtao Yu, ztyu@hotmail.com

Abstract: Paper cooperation network embodies expert topic similarity in an extent, thus, a novel method is proposed for expert topic analysis and extraction by combining paper cooperation network and topic model. In the method, we extract each paper' author information and construct an expert cooperation network. At the same time, by means of LDA model, a probabilistic topic model is also built to analyze papers' latent topics. Then, by making full use of the feature that adjacent nodes in the expert cooperation network share similar themes distribution, we makes a constraint on expert topic distribution in Gibbs sampling process of solving the probabilistic topic model. Experimental results on NIPS dataset show that the proposed method can effectively extract expert topics, and the expert paper cooperation network plays a very good supporting role on the extracting task.

Keywords: expert topic analysis, paper cooperation network, probabilistic topic model, Gibbs sampling, expert topic extraction

1. Introduction

Experts are very important resources in today's knowledge society. In all walks of life, experts, with broad professional knowledge, proficient skills and rich experiences, are urgently needed to review, guide, supervise and inspect all kinds of project approval, project implementation and project acceptance. From scientific research institution to social production departments, experts are also needed to organize a team, guide product development, and tackle key problems, so that these institutions and departments can improve work and production efficiency. Therefore, it is an important scientific problem to obtain the relevant experts from the vast network knowledge. Expert topic analysis and extraction plays an important role in expert search and expert recommendation.

Expert topic analysis is that extracts expert research areas and obtains expert topics from experts' homepages, experts' published papers, and experts' social networks. Existing topic analysis method is mainly divided into three categories. The first category is a topic analysis model based on Latent Semantic Analysis (LSA) [1]. In the category method, singular value decomposition in matrix theory is applied to divide a larger TF-IDF matrix into three smaller matrices to construct a new low-dimensional latent semantic space and find a simpler expression for the document. The second category is a probabilistic topic analysis model based on Probabilistic Latent Semantic Analysis (PLSA)

ACCEPTED MANUSCRIPT

[2-4]. It is a generative model. Similar to LSA, its goal is to find a transformation from lexical space to latent semantic space. PLSA achieves better results on topic extraction when compared with LSA. The third category is a topic analysis model based on Latent Dirichlet Allocation (LDA). This model is based on the PLSA and introduces a Dirichlet prior distribution on the basis of document-topic distribution and topic-term distribution, and constructs a topic analysis model based on full-probability LDA [5-7]. LDA model has been applied well in topic analysis. Many researches, such as parameter expansion [8-10], introducing context information [11-12], proposing topic model for specific task [13-15] and so on, have been conducted around LDA topic model. And this improves topic extraction effect on related tasks. In respect of expert topic analysis model, Steyvers et al. [16] proposed the Author-Topic (AT) model, that each author has a topic probability distribution. Later, on the basis of AT model, McCallum et al. [17] further proposed Author-Recipient-Topic (ART) model that using the directivity of e-mail interaction, takes the sender and receiver pairs as the decisive factor of a document topic probability distribution, and has a good performance on document topic extraction. Tang Jie et al. [18], in ArnetMiner system, proposed an Author-Conference-Topic (ACT) model based on the relationship between partners, which achieved very good results on expert topic extraction.

As the most direct research result of the experts, the thesis has rich expert topic information. For example, for a paper, the title, keywords, content etc. can adequately reflect its topics and research fields, and therefore, also reflect the expert topics and research areas of its authors. And this topic information can be effectively extracted by LDA model. For a same paper, there is a partnership among its multiple authors, and there is a guiding relationship between its corresponding author and first author. These mean that its authors share the same or similar research interests and have the same sharing field knowledge. The more papers several experts co-work, the more field knowledge the experts share, the more similar expert topic distribution the authors share. Just yet, the expert cooperative network can characterize these relationships between experts well. Existing expert thematic analysis methods commonly only use single expert information, such as the content and simple partnership of a paper, to make topic analysis. However, expert topic similarity distribution, which is characterized by experts' paper cooperative network, has played a positive role on expert topic analysis. Therefore, this paper discusses how to effectively use expert paper cooperation network to improve the effect of expert topic extraction. On the one hand, we extract each paper' author information and construct an experts' paper cooperation network. On the other hand, we make a topic analysis on papers and build a paper probabilistic topic model. Then, Gibbs sampling is used to solve the parameters of the model, and in this process, we makes a constraint through fusing the experts' paper cooperation network. Finally, the experts' topics are extracted. Figure 1 shows the flowchart of the proposed method.

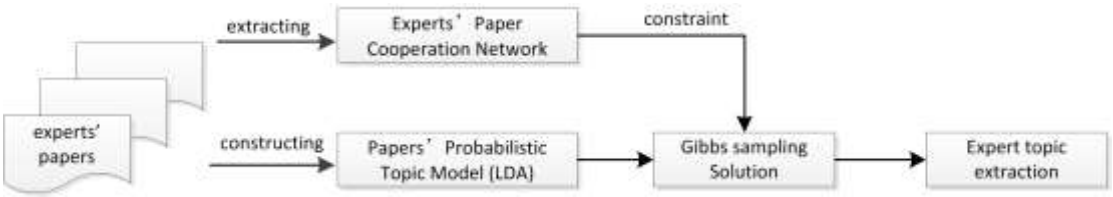


Fig. 1. Flowchart to the proposed method

The main contributions of the paper are: (1) Analyze the authors' cooperative relations on the papers to construct the experts' paper cooperation network. (2) Construct the probability topic model of the papers by using LDA. (3) Integrate the experts' paper cooperation network into the Gibbs sampling process of solving the papers' probabilistic topic model as a constraint. (4) Verify the effectiveness of the expert topic extraction by making experiments.

The rest of the paper is organized as follows: Section 2 describes expert cooperation network. Section 3 presents the probabilistic topic analysis for experts' papers. Section 4 constructs the expert topic model construction by integrating expert cooperation network and paper topic model. Experiments and analysis are provided in section 5. Then, conclude this paper in Section 6.

2. Probabilistic representation for expert cooperation network

2.1. Definition and formalization for expert cooperation network

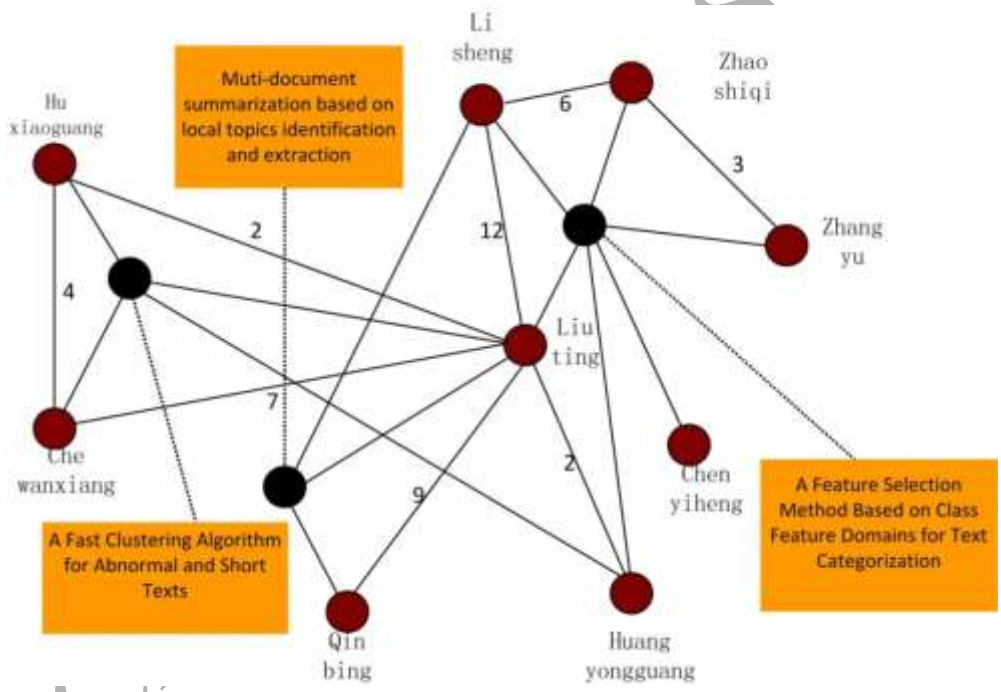


Fig. 2. An example of expert cooperative network

An article, published by an expert, tends to have multiple partners. An expert cooperation network can be build on the basis of paper partnership among experts. Collaborators in this network are often experts who share the same field knowledge. That is these experts have a very similar topic distribution. Based on the above analysis, expert cooperation network plays a supporting role on expert topic extraction. Figure 2 shows a simple expert cooperation network.

In Fig.1, a black solid dot represents a published article of experts, and all the red solid dots, which are connected to a black solid dot, represents all the experts who collaboratively published this paper. If there is more than once cooperation between the two experts, there is an undirected edge

to connect the two experts, such as the edge of Sheng Li-Ting Liu and that of Liu Ting-Qing Bin, and the times of cooperation between the two experts represent the weight of the edge. In order to better elaborate the process of building expert topic model by fusing expert cooperation network and probability topic model, we will formalize the expert papers and the expert cooperation network.

For an expert paper, all papers published by experts constitute a set of papers, that is tagged as C. Since each paper d in the set can be seen as a sequence of words $w_1 w_2 \ldots w_{|d|}$, the paper can be formally characterized as $d = \{w_1 w_2 \ldots w_{|d|}\}$.

For an expert cooperation network, which can be formally characterized as a graph $G = \langle V, E \rangle$. Wherein V is a set of all nodes in the graph, $v \in V$ represents a node in the node set V, and that is an expert. E is a set of all edges in the graph, which must conform with $(E_{v,d} \subset E) \cap (E_{v,\bar{v}} \subset E) \cap (E_{v,d} \cup E_{v,\bar{v}} = E)$. Among them, $E_{v,d}$ represents the set of edges that are formed expert v and his or her published papers d, which confirm with $\langle v, d \rangle \in E_{v,d}$. $E_{v,\bar{v}}$ represents the set of edges which are formed between expert v and expert $\bar{v}$ who cooperate more than once, which confirm with $\langle v, \bar{v} \rangle \in E_{v,\bar{v}}$. $c(v,\bar{v})$ characterizes the weight of edge $\langle v, \bar{v} \rangle$, which equals the number of cooperation between the two experts. The two type edges confirm with $\langle v, d \rangle = \langle d, v \rangle$ and $\langle v, \bar{v} \rangle = \langle \bar{v}, v \rangle$ respectively. Since edges in expert cooperation network are undirected, expert cooperation network G is an undirected graph network.

2.2. Probabilistic representation for expert cooperation network

According to expert cooperation network structure, there is not difficult to see that the adjacent expert nodes in the network have common characteristics. Namely adjacent experts must cooperate with each other more than once, and we believe that the experts who cooperate frequently should belong to same topics. Therefore, for the neighboring nodes in the expert cooperation network, we hope that the difference in their topic distribution is as small as possible. Based on the above analysis, the formal representation of the objective function of expert topic distribution in the expert cooperation network is shown as Eq.(1).

$$N(C,G) = \frac{1}{2} \sum_{\langle v,\bar{v} \rangle \in E} c(v,\bar{v}) \sum_{j=1}^{k} (f(\theta_j, v) - f(\theta_j, \bar{v}))^2 \qquad (1)$$

Herein, C denotes the set of experts' papers. G denotes the expert cooperation network. $c(v,\bar{v})$ denotes the weight of edge $\langle v, \bar{v} \rangle$, which means the times of cooperation between expert v and expert $\bar{v}$. $f(\theta_j, v)$ denotes the conditional probability $p(\theta_j \mid v)$ that expert v assigns to topic θ_j in a condition of given expert v. In the same way, $f(\theta_j, \bar{v})$ denotes the conditional probability $p(\theta_j \mid \bar{v})$ that expert $\bar{v}$ assigns to topic θ_j in a condition of given expert $\bar{v}$. Obviously, for an expert, the probability distribution on all topics conforms to $\sum_{i=1}^{k} p(\theta_i \mid v) = 1$. $\sum_{j=1}^{k} (f(\theta_j, v) - f(\theta_j, \bar{v}))^2$ denotes the sum of the differences between the probability distribution that expert v shares on k topics and the probability distribution that his adjacent expert $\bar{v}$ shares on k topics. Obviously, in order to make the differences between the topic distributions that adjacent nodes share as small as possible, we only need to minimize the objective function $N(C,G)$.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

3. Probabilistic topic Analysis for experts' papers

Expert topics are fully reflected in his papers. A paper perfectly conveys experts' research areas and experts' research contents, which imply abundant expert topic information. We analyze the process of writing a paper by experts: At the beginning, experts will firstly determine a topic for the paper to be written. On the condition of the determined topic, they will organize related words with the topic to compose the paper. By the time all the words are written, a paper will have been finished. This is in full compliance with the representation of LDA model. In this paper, therefore, LDA topic model, which has been proven to have excellent performance on topic extraction, is chosen to make topic modeling for the experts' paper set C.

LDA topic model introduces topic as a potential variable, and regards each document d as the probability distribution θ that the document shares on several topics z, and regards each topic z_m as the probability distribution φ that the topic shares on words w. By introducing Dirichlet prior distribution α and β into these two distributions θ and φ respectively, a three layer Bayesian probability model of "document-topic-word" is constructed. The model parameters are estimated by fitting the training data, and the generation process of document d is simulated to extract implicit topics of the document. The Bayesian network diagram of the LDA model is shown in Figure 3

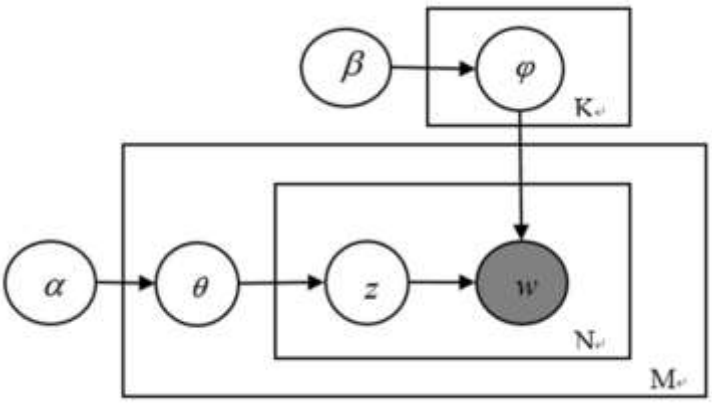


Fig.3. LDA diagram model representation

In Figure 2, the hollow circles represent implicit variables, including α, β, θ, φ, z. The solid circle represents the observable variable W, which represents a word in document d. Directed edges represent conditional probability dependence. The boxes mean repetitions, and the letters in the right bottom of the box denote the number of repetitions. The *N* means that repeatedly generating *N* words to form a document, the *M* means that repeatedly generating *M* times to generate a collection C of *M* documents. For each document d in the collection C, its generation process is as follows [19-20]:

Firstly, generate the document length according to the Poisson distribution $N \sim Poission(\xi)$;

Generate the document-topic distribution $\theta \sim Dir(\alpha)$ by using the Dirichlet distribution with the parameters α;

Using the Dirichlet distribution with the parameters β to generate the topic-word distribution $\varphi \sim Dir(\beta)$;

ACCEPTED MANUSCRIPT

For each word w_n in the document:

Generate a topic Z_n from the distribution $Multinomial(\theta)$ that the document shares on the topics;

Generate a word w_n from the distribution $Multinomial(\varphi)$ that the topic z_n shares on the words.

In the process of generating the document, N is a scalar, which represents the length of the generated document. ξ is a scalar that represents a parameter of the Poisson distribution. θ is a K dimensional vector which satisfies the multinomial distribution and generates document topic distribution. α is a K dimensional vector, representing the parameters of de Lickley distribution. φ is a N dimensional vector which satisfies the multinomial distribution and represent the topic word distribution of the topic generating. β is an N dimensional vector, representing the parameters of de Lickley distribution. w_n is an N dimensional vector, which represents the word of the n location to generate the document. If w_n is a word v and satisfies integer $1 \le v \le N$, the v position element of w_n is 1, the other locations elements are 0. z_n is a K dimensional vector that represents the subject topic of the n position word of the generating document. If z_n is a topic t and satisfies integer $1 \le t \le K$, the t position element of z_n is 1, the other locations elements are 0. The K represents the number of topics, and N represents the total number of words contained in the document collection.

Based on the above analysis, the joint probability distribution of the paper-topic and the topic-topic-words can be characterized by the formula (2).

$$p(w_m, z_m, \theta_m, \varphi \mid \alpha, \beta) = \left\{ \left[\prod_{n=1}^{N_m} p(z_{m,n} \mid \theta_m) p(w_{m,n} \mid \varphi_{z_{m,n}}) \right] p(\theta_m \mid \alpha) \right\} p(\varphi \mid \beta) \quad \ldots\ldots(2)$$

Among them, $p(z_{m,n} \mid \theta_m)$ characterizes the probability of generating the topic $z_{m,n}$ by the condition that the paper-topic distribution θ_m is known. $p(w_{m,n} \mid \varphi_{z_{m,n}})$ characterizes the probability of generating the topic word $w_{m,n}$ by the condition that the topic-topic-words distribution $\varphi_{z_{m,n}}$ is known. N_m characterizes the number of words in the entire paper. The parameters $p(\theta_m \mid \alpha)$ and $p(\varphi \mid \beta)$ are the posterior probability respectively. Obviously, the parentheses part in formula (2) describes the generating process of the words, and the braces part describes the generating process of the whole thesis. Since α is Dirichlet prior distribution θ_m of the paper-topic, β is the Dirichlet prior distribution φ of topic-topic-words, equation (2) can simplified as shown in formula (3) joint probability distribution $p(\vec{w}, \vec{z})$, represented by the Model $T(C)$.

$$T(C) = p(\vec{w}, \vec{z} \mid \vec{\alpha}, \vec{\beta}) = p(\vec{w} \mid \vec{z}, \vec{\beta}) p(\vec{z} \mid \vec{\alpha}) \quad \ldots\ldots (3)$$

Among them, the first factor $p(\vec{w} \mid \vec{z}, \vec{\beta})$ expresses the process of sampling words basing on the prior distribution parameter β of determined topic $\vec{z}$ and words distribution. The second factor $p(\vec{z} \mid \vec{\alpha})$ is the process of sampling the topic according to the prior distribution parameter α of topic distribution. The sampling process of words and the sampling process of topic are independent of each other, and the model parameters can be estimated by the Gibbs sampling method finally.

4. Expert topic model construction by integrating expert cooperation network and paper topic model

ACCEPTED MANUSCRIPT

4.1. The idea and derivation of expert topic model

Expert collaboration network plays a very important supporting role for the expert topic clustering. For example, two experts who often write a paper together are likely to be experts in the same research area. From the perspective of the topic analysis, they likely have the same topic distribution. Thus, though fusing the expert cooperation network and the probabilistic topic model, the expert topic extraction model is constructed. In the expert cooperation network, the difference of topic distribution between adjacent experts is little, we can use this characteristic to make a constraint on expert topic clustering. so we can improve the performance of expert topic clustering. According to the analysis on the expert cooperation network in the second section, we can know that the objective function of the expert topic distribution is formalized as $N(C,G)$, as shown in the formula (1). According to the analysis on the probabilistic topic model in the third section, we can know that the joint probability distribution of the paper-topic and the topic-topic-words is formalized as $T(C)$, the concrete formula as (3). Therefore, the constructed expert topic model fusing expert collaboration network and probabilistic topic model can be formalized as $M(C,G)$, and the model is shown in the formula (4).

$$\begin{aligned} M(C,G) &= (1-\pi)[-T(C)] + \pi N(C,G) \\ &= (1-\pi)[-p(\vec{w}\,|\,\vec{z},\vec{\beta})\,p(\vec{z}\,|\,\vec{\alpha})] + \frac{\pi}{2}\sum_{\langle v,\bar{v}\rangle\in E} c(v,\bar{v})\sum_{j=1}^{k}(f(\theta_j,v)-f(\theta_j,\bar{v}))^2 \end{aligned} \quad \ldots\ldots(4)$$

Among them, parameter $\pi \in [0,1]$ is introduced to balance the interaction between probabilistic topic model $T(C)$ and topic distribution objective function $N(C,G)$ in the process of topic clutering. When the balance factor π is taken 0, the expert topic model $M(C,G)$ is reduced to $[-T(C)]$. It means that the result of the expert's topic extraction is only determined by the solution $T(C)$ to the probabilistic topic model. When the balance factor π is taken 1, the expert topic model $M(C,G)$ is reduced to $N(C,G)$. In other words, the subject extraction problem is transformed into the optimization problem of the objective function $N(C,G)$. In a more general situation, when the balance factor $\pi \in (0,1)$, namely $0<\pi<1$, the extraction of the expert topic model $M(C,G)$ is affected by the probabilistic topic model $T(C)$ and the objective function $N(C,G)$ of the topic distribution in the expert cooperation network. As shown in formula (4), probabilistic topic model front with a minus sign in the expert topic model. Through the second section of the analysis of the network of experts can be known, we can know that our goal is to minimize the objective function in order to make the difference of the adjacent nodes in the topic distribution as little as possible. For the probabilistic model of the first half of the model, the model is solved by using Gibbs sampling. Gibbs sampling theory based on LDA tells us that Gibbs sampling is proportional to the state of posterior probability. That is, the probability of the posterior probability is high, and the probability of convergence to the state of Gibbs sampling is also high. Thus, the goal of probabilistic topic model $T(C)$ is to maximize the posterior probability, and it is difficult to unify the posterior probabilistic of minimizing the objective function $N(C,G)$ and maximizing the probabilistic topic model $T(C)$. So in order to solve the unified model, probabilistic topic model $T(C)$ is coupled with a minus sign in the construction of experts topic model $M(C,G)$ to make the model solving process to equal to minimization function $M(C,G)$. When $M(C,G)$ reaches the minimum, the probability distribution of expert-topic and topic-word will be obtained.

ACCEPTED MANUSCRIPT

4.2. Expert topic model solving and expert topic extraction

Gibbs sampling is a special case of Markov Chain Monte Carlo (MCMC) algorithm. When it is hard to directly sampling, Gibbs sampling can obtain a series of sample sequences from the specified multivariate probability distribution, and these sample sequences can be used to approximately estimate the joint probability for all variables and the marginal probability for a single variable. Gibbs sampling is mainly applied to the situation that the joint probability can't be obtained directly or is hard to directly sample, but the conditional probability distribution of each variable is known and easy to sample. In addition, MCMC algorithm is based on the convergence theorem of Markov chains. Thus, Gibbs sampling is more effective than EM algorithm. We, therefore, apply Gibbs sampling to estimate the LDA model parameters so as to implement the solving of the expert topic model. The solving process of the model is equivalent to the minimum function $M(C,G)$, which is composed of the probabilistic topic model $T(C)$ and t the objective function $N(C,G)$ of the topic distribution in the expert cooperation network. We first consider obtaining the optimal solution of the probabilistic topic model $T(C)$ by the means of Gibbs sampling and judge whether the topic distribution of the target function of the expert cooperation network is close to the minimum under this condition. Obviously, we want $\sum_{j=1}^{k}(f(\theta_j,v)-f(\theta_j,\bar{v}))^2$ to be nearly 0. That is, in the expert cooperation network, the neighboring experts have the same topic distribution as much as possible. As a result, the objective function $N(C,G)$ can be quantified as close to 0. If $N(C,G)$ tends to 0, the model achieves optimal; otherwise, the parameters of the model are reinitialized, and parameters are tuned using Gibbs sampling until the global optimal solution is obtained. The solution of the model is illustrated in Figure 4.

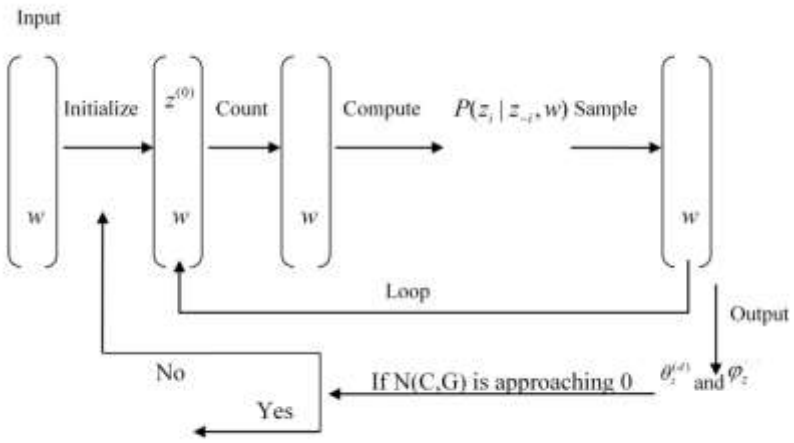


Figure 4 expert topic model solving process

The following is the necessary derivation of the Gibbs sampling procedure for a probabilistic topic model $T(C)$. According to the analysis of the probabilistic topic model in section third, we need to solve the joint probability distribution $p(\vec{w},\vec{z})$ of the formula (3). Due to the sampling process and the topic of the sampling process independent of each other in the formula (3), two processes can be separated in the Gibbs sampling process.

ACCEPTED MANUSCRIPT

For the sampling process $p(\vec{w}|\vec{z},\vec{\beta})$ of the word, the word distribution can be generated based on the determined topic $\vec{z}$ and the word distribution sampled from the prior distribution β, as shown in the formula (5).

$$\begin{aligned} p(\vec{w}|\vec{z},\varphi) &= \sum_{i=1}^{w} p(w_i|z_i) = \sum_{i=1}^{w} \varphi_{z_i,w_i} \\ &= \prod_{k=1}^{K} \prod_{\{i:z_i=k\}}^{K} p(w_i=t|z_i=k) = \prod_{k=1}^{K}\prod_{t=1}^{V} \varphi_{k,t}^{n_k^{(t)}} \end{aligned} \tag{5}$$

Among them, $n_k^{(t)}$ is the number of times the word t appeared in the topic k.

The distribution of the target $p(\vec{w}|\vec{z},\vec{\beta})$ is solved by integrating the word distribution φ formula (5), and the formula is shown in (6).

$$p(\vec{w}|\vec{z},\vec{\beta}) = \int p(\vec{w},\vec{z}|\varphi)p(\varphi|\vec{\beta})d_\varphi = \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})} \tag{6}$$

The sampling procedure $p(\vec{z}|\vec{\alpha})$ for the subject matter can be produced according to the prior distribution parameter α of the subject distribution, as shown in (7).

$$\begin{aligned} p(\vec{z}|\theta) &= \prod_{i=1}^{W} p(z_i|d_i) \\ &= \prod_{m=1}^{M}\prod_{k=1}^{K} p(z_i=k|d_i=m) = \prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{m,k}^{n_m^{(k)}} \end{aligned} \tag{7}$$

Among them, d_i is the document attached to the word i, and $n_m^{(k)}$ represents the number of times that the topic k appears in the document m.

In the same way, the distribution of the target $p(\vec{z}|\vec{\alpha})$ is solved by integrating the word distribution θ formula (7), and the formula is shown in (8).

$$p(\vec{z}|\vec{\alpha}) = \int p(\vec{z}|\theta)p(\theta|\vec{\alpha})d_\theta = \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \tag{8}$$

The joint probability distribution $p(\vec{w},\vec{z})$ in the probabilistic topic model $T(C)$ can be obtained by combining formula (6) and the formula (8) , as shown in the formula (9).

$$T(C) = p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha}) = \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \tag{9}$$

After the joint probability distribution of the probability subject model $T(C)$ is obtained, the Gibbs sampling formula (10) is obtained according to the Bayesian rule and de Lickley's prior distribution.

$$p(z_i=k|\vec{z}_{\neg i},w) \propto \frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)}+\alpha_k\right)} \cdot \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(t)}+\beta_t\right)} \tag{10}$$

ACCEPTED MANUSCRIPT

Among them, $\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)}+\alpha_k\right)}$ is characterized by the probability distribution of the expert-topic, and $\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(t)}+\beta_t\right)}$ is characterized by the probability distribution of the topic-topic-words.

Through repeated iteration of the Gibbs sampling process, the sampling results are finally stable. The posterior distribution of the topic in the published paper collection and the posterior distribution of the word in each topic can be calculated by each expert. The final solution of distribution expectations of the expert-topic and the distribution expectations of the topic-topic-words are as shown respectively as the formula (11) and the formula (12).

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k} \quad (11)$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \quad (12)$$

Among them, $n_k^{(t)}$ is the number of times that the word t is assigned to the topic k, and $n_m^{(k)}$ is the number of words assigned to the topic k in the document m. According to experience, we set the two parameters of the LDA model, $\alpha=50/K$ (K is the number of topics) and $\beta=0.01$, respectively[21]. After solving the distribution expectations $\theta_{m,k}$ of experts-topic and the distribution expectations $\varphi_{k,t}$ of the topic-topic-words, $\theta_{m,k}$ is brought into the formula (1) to determine whether the topic distribution of the target function is close to 0 in the expert cooperation network. If it is close to 0, the model is to achieve the best overall and it is the final results of the experts topic model at this time; otherwise, the parameters of model are reinitialized , and parameters are tuned using Gibbs sampling until the objective function of the topic distribution in expert cooperation network is close to 0. At this point to solve the $\theta_{m,k}$ and $\varphi_{k,t}$ for the final results of the expert topic model.

5. Experiment and analysis

5.1. Experimental data set

In this paper, we use NIPS data set [22] to verify the proposed expert topic model, which fuses expert cooperation network and probabilistic topic model, to verify the topic clustering effect that the model can obtain on expert topic extraction task. The data set contains all conference papers that were published in Annual Conference on Neural Information Processing Systems from 1987 to 1999. Each paper contains title, authors, key words, publication date and text content. Furthermore, the author information is extracted from each paper and saved in a separate file, and this provides a convenient and effective way for us to construct expert cooperation network. In order to extract the expert topic well, in preprocessing stage, we filter stop words and phrases with the help of a stop word list, such as "the", "a" and so on, and reserve domain vocabularies, such as "SVM", "Hmm",

ACCEPTED MANUSCRIPT

"KNN", "CRF", "AI" and so on. When analyzing the dataset, we found some words that occur less than five times. Most of these words are come from OCR error when we use PDF2TXT tool to convert PDF into TXT. Therefore, word frequency statistics are calculated on the data set through the word frequency statistical tool, and the vocabularies, whose frequency is less than five times, are filtered out. After the dataset is preprocessed by the above processed, the number of papers, the number of experts and the total number of words are counted, and the results are shown in Table 1. For expert disambiguation problem, we adopt Chinese expert disambiguation method based on semi-supervised graph clustering proposed by Jiang [23].

Table.1. Data set

the number of papers	the number of experts	the total number of words
1740	2037	13649

5.2. Evaluation

In this paper, we select the standard evaluation criteria (perplexity) [5] as the evaluation index of the model. By calculating the perplexity of a given test set, the ability of that the model generates text can be evaluated. The lower the perplexity is, the better the text generalization ability of the model is [24]. The model perplexity is calculated as follows (13).

$$\text{perplexity}(\mathrm{D}_{\text{test}})=\exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{13}$$

Among them, D_{test} denotes the test set, $\sum_{d=1}^{M}N_d$ denotes the number of words on the test set, and $p(w_d)$ denotes the probability of that document d generates word w on test set.

5.3. Experiment design and result analysis

In order to verify the effectiveness of the expert topic model proposed in this paper, we design three experiments. The first experiment is an intuitive presentation of expert topics which are extracted by the proposed model. The second experiment compares the perplexity generated by the proposed model and that one generated by the traditional LDA model, when topic number is same and iteration times of Gibbs sampling is different. The third experiment compares the perplexity generated by the proposed model and that one generated by the traditional LDA model, when topic number is different and iteration times of Gibbs sampling is same. In addition, the time complexity of the proposed method is $O(C*K*N)$. Herein, C denotes iterative number, K denotes topic number, and N denotes word number. So, the computational efficiency is much the same with the traditional LDA model.

The first experiment: The clustering number is set to 15, and the Gibbs sampling iteration number is set to 1000. Expert topic extraction is carried out on the preprocessed NIPS data sets by the expert topic model combining the expert collaboration network and the probabilistic topic model. Due to the extracted topics is more, and the extracted topic words, which are assigned to a topic, are more, these all are not convenient to display in this paper. Therefore, we only select three representative expert topic and the top 10 topic words with maximum probability for each topic to display. Part of the extracted topic results are shown in table 2.

ACCEPTED MANUSCRIPT

Table.2. Part of the extracted topic results

The topic and lexical entry	Probability	The topic and lexical entry	Probability	The topic and lexical entry	Probability
Topic3	0.3274	**Topic10**	0.1753	**Topic14**	0.0526
speech	0.08234	system	0.09245	support	0.08270
recognition	0.07456	control	0.08526	likelihood	0.05408
context	0.04239	function	0.06944	recognition	0.04756
word	0.03561	basic	0.06413	features	0.03842
words	0.02749	vol	0.04237	feature	0.02134
hmm	0.02148	feed forward	0.03781	system	0.02077
training	0.01473	actions	0.02459	distribution	0.01473
speaker	0.01422	paper	0.01873	face	0.01246
system	0.01267	em	0.01698	mixture	0.01046
feature	0.01154	work	0.01264	context	0.00867

Table 2 shows the topic extraction results of the speech recognition domain expert De-Mori. The probability of its number 3 topic is 0.3274, which is the one of maximal probability from 15 extracted topics. By observing the Topic3 in Table 2, it is not difficult to find that the first 10 topics of the maximum probability closely relate to the field of speech recognition on the Topic3, which is consistent with the fact that De-Mori is an expert in the field of speech recognition. Experimental results show that the model proposed in this paper has achieved good results in expert topic extraction.

The second experiment: The clustering number is set to 15. The proposed method and traditional LDA model are respectively used to extract the expert topics, and the perplexity is calculated in different Gibbs sampling iterations under two models. The experimental results are as shown in Figure 5.

ACCEPTED MANUSCRIPT

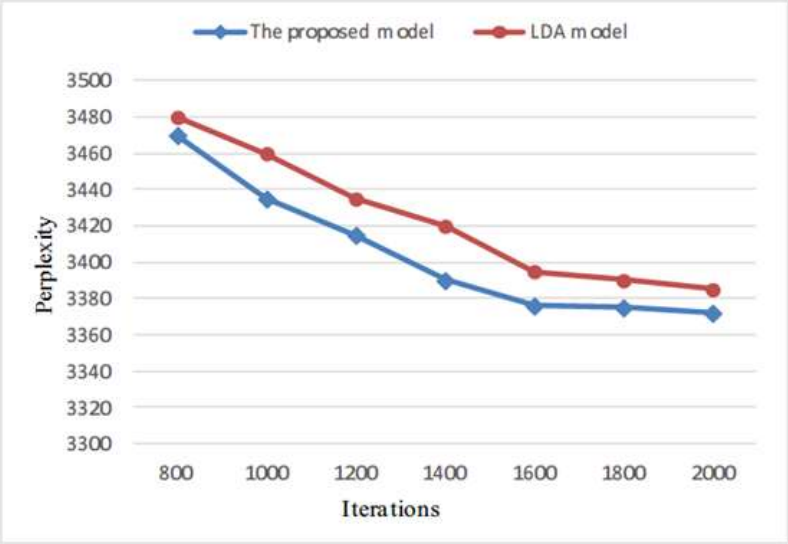


Fig. 5. Comparison of the perplexity of the two models

By analyzing the experimental results of experiments two, it is found that the perplexities of the expert topic model fusing the expert cooperation network and the probabilistic topic model presented in this paper and the LDA model are decreased with the increase of Gibbs sampling iteration times. Finally, the tendency is stable, but the perplexity of the model proposed in this paper is always lower than the LDA model. The experimental results show that the generalization of the model is effectively improved through the fusion of expert cooperation network and the probabilistic topic model.

The third experiment: The numbers of topics are set to 5, 10, 15, 20, and 25, the times of Gibbs sampling iteration number is set to 1000. The proposed method is respectively used to extract the expert topics under different number of topics, and the perplexity of each model is calculated. The results are shown in figure 6.

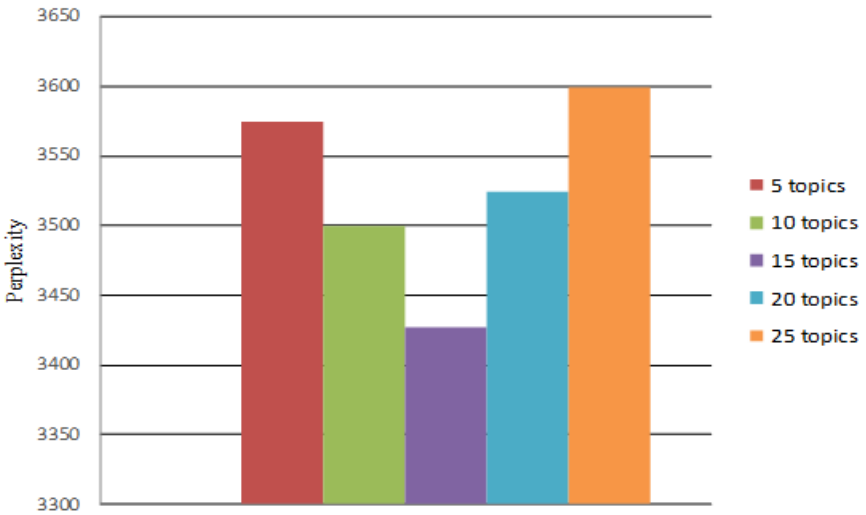


Fig. 6. Comparison of perplexities under different number of topics

ACCEPTED MANUSCRIPT

Through the analysis of experiment three, it is found that the perplexity of the model first decreases and then increases with the increase of the number of topics. When the number of topics is 15, the perplexity of model is lowest. The experimental results show that there is an optimal number of topics for a certain set of data, which makes the generalizability of model better.

6. Conclusion

To a certain extent, expert cooperation relationship and paper topic can represent experts' subjects. In this paper, we integrate expert cooperation network and probabilistic topic model to make expert topic extraction. Based on the analysis of paper topics, the cooperation relationship of experts' papers is made full use to improve the accuracy of expert topic analysis. Experiments also demonstrate that the proposed method is effective, and that the expert cooperation network has a good supporting effect on expert topic extraction. Among experts, besides the explicit cooperative relationship on papers, there are many other content-related features, such as topic relevance and citation correlation among papers. The further research works will focus on how to effectively utilize more implicit relationships to enhance the effectiveness of expert topic analysis.

Acknowledgments

This paper is supported by National Nature Science Foundation (Nos.61175068, 61472168), the Key Project of Yunnan Nature Science Foundation (No.2013FA030), and the Science and technology innovation talents fund project of Ministry of Science and Technology (No.2014HE001).

References

[1] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman. Deerwester, Scott, Susan T. Dumais. Indexing by latent semantic analysis. *Journal of the American society for information science*, 1990, 41(6): 391-391.

[2] T. Hofmann. Probabilistic latent semantic analysis. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1999, pp. 289-296.

[3] T. Hofmann. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1999, pp. 50-57.

[4] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 2001, 42(1-2): 177-196.

[5] D.M. Blei, A.Y. Ng, M.I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 2003 3(2003): 993-1022.

[6] T.L. Griffiths, M. Steyvers. Finding scientific topics. In: *Proceedings of the National academy of Science of the United States of America*, 101 Suppl 1, 2004, pp. 5228-5235.

[7] S.X. Gao, Z.T. Yu, L.B. Shi, X. Yan, H.X. Song. Review expert collaborative recommendation algorithm based on topic relationship. *IEEE/CAA Journal of Automatica Sinica*, 2015, 2(4):403-411.

[8] D.M. Blei, J.D. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 2005, 18(2005):113-120.

ACCEPTED MANUSCRIPT

[9] J. Zhang, J. Tang, H.L. Zhuang, C.W. Leung, J. Li. Role-aware conformity influence modeling and analysis in social networks. In: *Proceedings of the National Conference on Artificial Intelligence*, AI Access Foundation, 2014, vol.2, pp. 958-965. 2014.

[10] S.X. Gao, Z.T. Yu, W.X. Long, W. Ding, C.T. Yan. Chinese-Vietnamese Bilingual News Event Storyline Analysis Based on Words Co-occurrence Distribution. *Journal of Chinese Information Processing*. 2015, 29(6):90-97. doi: 10.3969/j.issn.1003-0077.2015.06.013. (In Chinese)

[11] T.L. Griffiths, M. Steyvers, D.M. Blei, and J.B. Tenenbaum. Integrating topics and syntax. In: *Proccedings fo 18th Annual Conference on Neural Information Processing Systems, NIPS 2004*, Neural information processing systems foundation, 2005, pp. 537-544.

[12] X.R. Wang, A. McCallum, X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proccedings of *Seventh IEEE International Conference on Data Mining, ICDM 2007*, IEEE, 2007, pp. 697-702.

[13] D.M. Blei, J.D. Mcauliffe. Supervised topic models. In: *Advances in neural information processing systems*, 2010, 3(2010):121-128.

[14] L. Hou, J.Z. Li, Z.C. Wang, J. Tang, P. Zhang, R.B. Yang, Q. Zheng. NewsMiner: Multifaceted news analysis for event search. *Knowledge-Based Systems*, 2015, 76 (2015):17-29.

[15] S.X. Gao, Z.T. Yu, J.J. Zou, M. Xiao, J.Y. Guo. Restricted Domain Question-Answering Text Retrieval Method Based on Supervised Latent Dirichlet Allocation Model. *Sensor Letters*, 2014, 12(2):380-385.

[16] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths. Probabilistic author-topic models for information discovery. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2014, pp. 306-315.

[17] A. McCallum, A. Corrada-Emmanuel, X.R. Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *Emmanuel*, 2013, 2013(2013): 1-10.

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

[19] J. Zhang. Probabilistic graphical modeling on heterogeneous social networks and applied to Information retrieval. Master's thesis, Beijing: Tsinghua University, 2009. (In Chinese)

[20] K. Cui. The research and implementation of topic evolution based on LDA. PhD dissertation., Beijing: National University of Defense Technology, 2010. (In Chinese)

[21] G. Xu, H.F. Wang. The Development of Topic Models in Natural Language Processing. *Chinese Journal of Computers*, 2011, 34(8):1423-1436. doi: 10.3724/SP.J.1016.2011.01423. (In Chinese)

[22] Sam Roweis. NIPS dataset (2002). *URL http://www.cs.toronto.edu/~roweis/data.html*. (accessed 2016.6.10)

[23] J. Jiang, X. Yan, Z.T. Yu, J.Y. Guo, W. Tian. A Chinese expert disambiguation method based on semi-

ACCEPTED MANUSCRIPT

supervised graph clustering. *International Journal of Machine Learning and Cybernetics*, 2015, 6(2):197-204.

[24] J. Cao, Y.D. Zhang, J.T. Li, S. Tang. A Method of Adaptively Selecting Best LDA Model Based on Density. *Chinese Journal of Computers*, 2008, 31(10):1780-1787. doi: 10.3321/j.issn:0254-4164.2008.10.012 (In Chinese)

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Shengxiang Gao is currently a Ph.D. candidate at Kunming University of Science and Technology, Kunming, China. She is also a CCF member since 2013. She received her M.S. degree in Pattern Recognition and Intelligent System from Kunming University of Science and Technology in 2005. Her research interests are in nature language processing, social computing, information retrieval and machine translation.

Xian Li is currently a M.S. candidate at Kunming University of Science and Technology. His research interests are in natural language processing, social computing and information retrieval.

Zhengtao Yu is currently a professor and Ph. D supervisor at the School of Information Engineering and Automation, and he is also the chairman of Key Laboratory of Intelligent

Information Processing, Kunming University of Science and Technology, Kunming, China. He received his Ph.D. degree in computer application technology from Beijing Institute of Technology, Beijing, China, in 2005. His main research interests are in natural language processing, social computing, information retrieval and machine translation.

Yu Qin received his M.S. degree in Technology of Computer Application from Kunming University of Science and Technology in 2015. His research interests are in natural language processing and information retrieval.

Yang Zhang is currently a M.S. candidate at Kunming University of Science and Technology. His research interests are in natural language processing, social computing, information retrieval and machine translation.

ACCEPTED MANUSCRIPT