International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.111-120
http://dx.doi.org/10.14257/ijdta.2016.9.12.11

Discovery of Subject of Science and Technology Policy based on LDA Model

Jianmin Wang [1], Shaobo Li[2], Tongrang Fan[3*] and Hui Wang[4]

School of Information Science and Technology, Shijiazhuang Tiedao University, China
Blacklion75@126.com[1], Fantr2009@126.com[2]*

Abstract

With the continuous increase of science and technology policy, how to find valuable information from the massive scientific and technological policies has becoming an urgent problem to be solved. Thus, this paper proposes a subject discovery method for large scale science and technology policy set. Based on LDA subject model for science and technology policy document subject modeling, this approach extracts time and geographical labels of science and technology policy, computes intensity of subjects under different time and geographical conditions, and obtains important subjects and analysis of theme change trend of the intensity of subjects under condition constraints. Experimental results demonstrate that this method can excavate and analyze the subjects form large scale science and technology policies quickly and effectively.

Keywords*: science and technology policy, LDA model, subject discovery, subject intensity.*

1. Introduction

Science and technology policy is the basic action standard of a country for the realization of scientific and technological tasks in a certain historical period. With the continuous development of science and technology, corresponded science and technology policy are constantly introduced resulting in the number of science and technology policy is increasing year by year. It has important significance that mining valuable information from the massive science and technology policy, and helping intelligence researchers to analysis of the current situation and policy and grasps the dynamic development of science and technology.

The current science and technology policy analysis methods are usually using subject discovery approach. Subject discovery is the process of quickly obtaining subjects from large-scale text messages using some ways. Subjects mining from texts not only can quickly grasp the core idea of the document, but also achieve effective dimension reduction of high dimensional data. The current subject discovery methods can be divided into two classes: clustering based methods and topic model based methods.

RM Aliguliyev[1] computes sentence dissimilarity using sentences as units of documents, clusters sentences via genetic algorithm, and selects sentences with larger weights in the above clustering as document subjects. Ang Zhao[2] et al. represent the original documents as graph structure using similarity between word and word, convert the document subject discovery as problem of graph segmentation, and use the spectral clustering algorithm to obtain the word clustering results, where each cluster is represented as a document subject.

Blei et al. propose[3] Latent Dirichlet Allocation (LDA) model in 2003. This model infers some subjects from texts by simulation of the text generation process. LDA subject model can efficiently mining hidden subject[4] information from large-scale document

ISSN: 2005-4270 IJDTA

collections. Thus, using LDA model has become the most commonly used method for subject modeling of text collection[5].

Facing on the subject discovery problem of large-scale science and technology policies, this paper proposes a science and technology policy set subject modeling method based on LDA model, introduce the concept of subject intensity[6], uses released times and scopes of implementation information of science and technology policies on the basis of computing subject intensity, and realizes the analysis of each subject under different conditions of time and region intensity change trend.

2. Related Knowledge

2.1 Construction of Subject Library of Science and Technology Policy

Subjects are the smallest information units that can represent ideas of original papers. Construction of specific library of science and technology policy domain is the basis for realizing important scientific subject discovery. In the actual indexing work, the scopes of subject discovery of science and technology policy are controlled by the limited special fields. Thus, we construct the standard "science and technology policy subject library" based on "public document subject table" issued by the Chinese Academy of Sciences and "keywords table of science and technology policy" which is indexed by specialist of policy researcher from a large number of science and technology policies. The above subject library contains science and technology management, international exchanges and cooperation, talent education in 15 categories and 2000 key words, and increases the number of subjects covered by each topic in the context of maintaining the number of topics in the subject area. The newly constructed subject library covers the basic science and technology policy in the common words, and provides a professional basis for the follow-up policy subject mining work. Table 1 describes some of the science and technology policy subject libraries, which include the subject areas and their corresponded subject words.

Table 1. Some Science and Technology Subject Words

Number	Subject area	Subject words
1	science and technology management	plan, strategy, innovation, index, science and technology…
2	scientific and technological cooperation	cooperation, passports, guests, visits, immigration...
3	scientific and technical personnel	university, academician, professional, educational background, degree...
4	Scientific and technological achievements and intellectual property rights	patents, results, registration, certification, copyright...
5	Rural science and technology	agricultural products, new rural areas, agriculture, grain...
6	Tax incentives	Taxation, budgeting, auditing, finance, accounting.....

2.2 Introduction of LDA Model

LDA subject model[7] is a kind of unsupervised machine learning method, which is used to identify the potential subject information in large scale document collection or corpus. LDA subject model is a three layer Bayesian probability model, which is composed of the text layer, the theme layer and the feature layer demonstrated in Figure 1. The LDA model assumes that the text is composed of a number of hidden subjects that is composed of a number of characteristic words, where the text to the subject follows the

deLickley distribution; subject to the characteristic words obey the polynomial distribution.

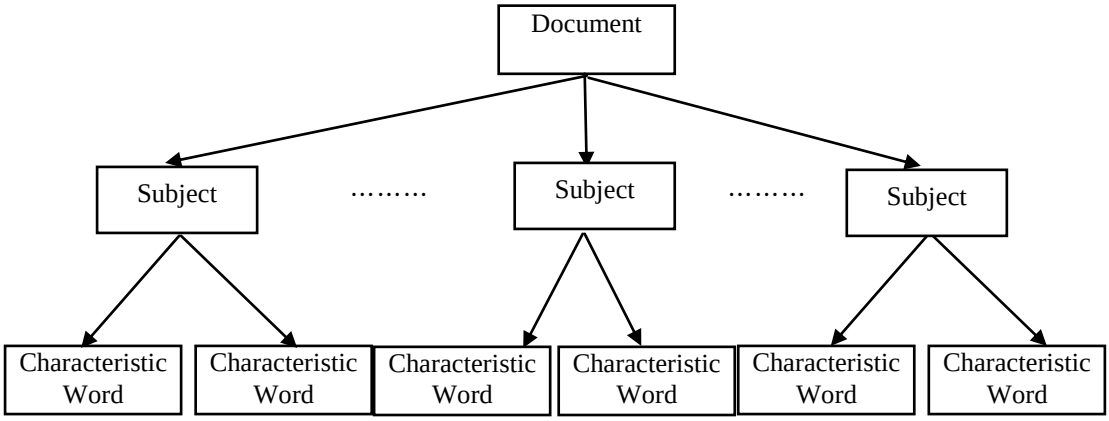

Figure 1. The Theme Layer and the Feature Layer Demonstrated of LDA Subject Model

Given document set $D=\{d_1,d_2,...d_m\}$ and potential subject set $Z=\{z_1,z_2,...z_t\}$, where $di=\{w_{i1},w_{i2},...w_{in}\}$ is the i-th document, m is the number of document, w_{ij} represents the j-th text in document d_i, n represents the text numbers in document d_i, and t is the number of subjects, the document generation process of LDA subject model is as follows:

1. Computing polynomial distribution of feature words and themes φ of feature words and subject based on Dirichlet priority distribution for the hidden subject I;

2. Obtaining the text number N of document based on Poisson distribution;

3. Calculating the probability distribution θ of the subject in each text;

4. Determining the feature word for each document of every document set;

4.1 Selecting a hidden subject z randomly from the subject probability distribution θ;

4.2 Choosing a feature word randomly from the polynomial distribution of subject z.

Where α represents the relative strength of the hidden subject from the document set, β is the probability distribution of the hidden subjects, θ represents the probability distribution of document subjects, and φ is the probability distribution of feature words in the hidden subjects.

3 Subject Discovery of Science and Technology Policies based on LDA Model

With the time and geographical differences, subjects included in the science and technology policies and the subject intensity will also be changed. It has important practical significance that analysis the evolution of the subject under different conditions for large-scale science and technology policy text collection to help policy researchers quickly finding the law of scientific and technological development. Thus, based on using LDA subject model for mining hidden subject information from science and technology policies, we use release time and the implementation of range label information of science and technology policies, introduce the concept of subject importance degree, and analysis the strength change according to the size of importance degree under different time and geographical conditions. This module is divided into three parts: text preprocessing, LDA topic modeling, and subject strength analysis.

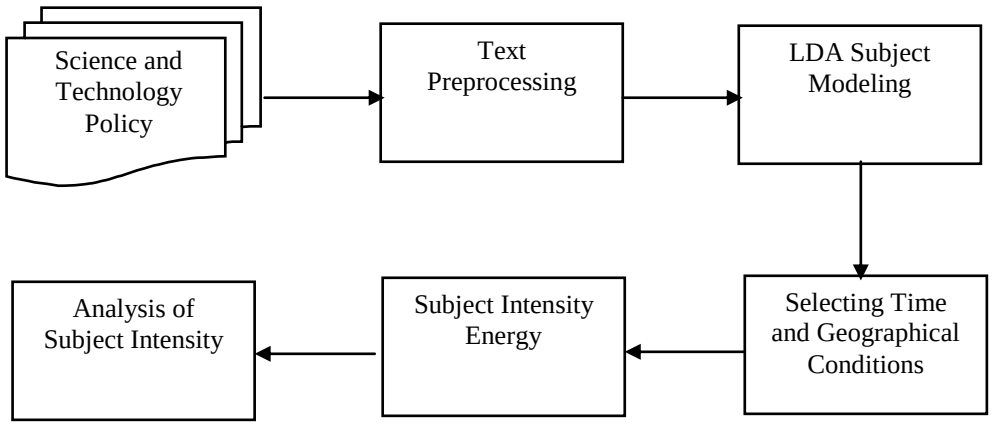

Figure 2. The Flow Chart of Subject Strength Analysis

3.1 Text Preprocessing

The pre-processing of texts of science and technology policy mainly includes four parts: format conversion, tag information extraction[8], word segmentation and remove stop words, feature selection.

1. Using obtained national and provincial science and technology policy in the text as the research object, this module converts the obtained HTML data into text format, and extract information of policy name, release time, and implementation scope label. The standard format data shown in table 2 will be stored in the database of scientific and technological policy, as a corpus of science and technology policy.

2. Contents of science and technology policy are carried out the Chinese word segmentation and stop words removal operations. In the word segmentation process, ICTCLAS segmentation tool of Chinese Academy of Sciences is use to achieve the text word segmentation, word frequency statistics and part of speech tagging operations. After that stop words are removed from the obtained word set for representing the policy texts to words set.

Table 2. Science and Technology Policy Saving Format

No.	Policy Name	Policy Content	Scope	Release Time
1	Beijing Patent Protection and Promotion Rules	Content	Beijing City	2013/09/27
2	Hebei Technology Market Regulations	Content	Hebei Province	2013/01/01

3. The purpose of feature selection is to calculate the weights of the words in the vocabulary, and choose the higher weights to construct the feature set of the text, so as to achieve the goal of dimension reduction. The traditional TF-IDF weighting algorithm only considers the frequency of words and ignores their meaning. Thus, it is prone to lose valuable features. According to the particularity of science and technology policy text, combined with "science and technology policy key words table", we improve the traditional TF-IDF method. We regard that if the processed words match with the subject in "science and technology policy key words table", those words should be given higher weights. Based on the above assumption, the improved weighted formula is defined as follows:

$$W(t) = \sum_{d \in D} (1+\lambda) \times TFIDF(d,t) \quad (1)$$

where λ is the subject factor, TFIDF(d, t) represent the weight of word t in document d.

We compute the weight of each word in the document, and set threshold δ. If $W(t_i) \geq \delta$, then choosing t_i as feature, otherwise deleting t_i from word set. At last, we obtain the feature set of text $d_i=\{(t_1,w_{i1}), (t_2,w_{i2})...(t_n,wi_n)\}$, where w_{ij} represents weight of word t_j in document d_i, and n represents the number of features.

3.2 LDA Subject Modeling

Based on the LDA model described in the section 2, the preprocessed text set is subject modeled. Since the LDA model requires the user to define the number of subjects before the implementation of the algorithm, and the choice of the number of topics has a great impact on the results of the LDA modeling, it is required to determine the optimal number of the subjects. In this paper, we use the evaluation criteria of perplexity to determine the optimal number of subjects, where evaluation criteria is based on the generalization ability of the model to automatically determine the optimal number of subjects suitable for the current LDA modeling. Generally speaking, the smaller value of the perplexity, the stronger the generalization ability of the model, that is, the number of the current subjects is the optimal solution.

The calculation equation is as follows:

$$\mathrm{pre}(D) = \exp\left| -\frac{\sum_{i=1}^{M} \left|\log(p(d_i))\right|}{\sum_{i=1}^{M} N_i} \right| \quad (2)$$

Where M is the number of texts in document set, N_i is the length of the i-th document d_i, and $p(d_i)$ is the probability of LDA model generating the document d_i.

In the process of subject modeling, subjects are extracted with different numbers of subjects, and their corresponding perplexities are also computed respectively. Based on the change trend of perplexities, the subject number with minimal perplexity is selected as the optimal subject number. At last we execute the LDA subject model to get the text collection corresponded document-subject matrix DT and subject-feature words matrix TW.

3.3 Subject Intensity Energy

Subject intensity is a measure of the criterion of subject strength. In this paper, we define the subject intensity $Sig(t_i)$, and analyze the change trend of subject intensity under different conditions based on computation of subject strength. Text is a mixed distribution of several subjects. The more information of subjects contained in texts, the more important of the subject to the text, based on this idea we compute the subject intensity $Sig(t_i)$ corresponded to document set.

Definition 1 subject intensity $Sig(t_i)$ of subject ti corresponded to document set D: Assuming $D=\{d_1,d_2,...d_m\}$ representing document set, $T=\{t_1,t_2,...t_n\}$ representing subject set obtained from modeling, t_i is the i-th subject in T, $d_j \in D$, $1 \leq i \leq N$, $1 \leq j \leq M$. $dj=(t_1,t_2,...t_n)$ is the subject probability distribution of the document d_j, the intensity of subject t_i corresponded to document set is defined as follow:

$$T(t_i, d_j) = P(t_i \mid d_j) = \frac{t_i}{\sum_{k=1}^{N} t_k} \quad (3)$$

$$I(t_i, d_j) = \log \frac{\sum_{k=1}^{D} t_i^k}{t_j^i} \quad (4)$$

$$Sig(t_i) = \frac{\sum_{k=1}^{D} T(t_i, d_j) \times I(t_i, d_j)}{M} \quad (5)$$

where M is the document number in D, $T(t_i, d_j)$ represents the intensity probability of the subject t_i in document d_j. If the value of $T(t_i, d_j)$ is larger, then the more important of ti to document d_j. $I(t_i, d_j)$ indicate subject t_i in d_j is common or rare, the larger value of $I(t_i, d_j)$, the more different of t_i and d_j, and the more not important of t_i and d_j. Combined with $T(t_i, d_j)$ and $I(t_i, d_j)$, we can obtain the final subject intensity $Sig(t_i)$. The larger value of $Sig(t_i)$ represents that subject t_i contain more information, and the greater intensity of the t_i in the document set.

4 Experiments and Analysis

4.1 Experimental Data

Experimental data is obtained by Crawler technology from public released policy documents of nation and server provinces, a total of 7500. The implementation scopes of those policies include nation, Hebei Province, Henan Province, Guangdong Province, Tianjin, Beijing and other regions in the time interval between 2012 and 2015. We first extract released time and implementation scope of each policy document respectively, and then convert to pure corpus suitable to experiments by text preprocessing of policy contents.

4.2 Experimental Steps

Using the Gibbs sampling method for parameter estimation in the LDA modeling process, for the prior super parameters selection of α and β, it is defined that $\alpha = 50/k$, $\beta = 0.01$, where k is the subject number before LDA modeling. Since the defect of LDA model needing predefine the subject number, the perplexity criterion is used to determine the optimal number of subject number k with the following steps in detail: different values of k 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 are used for LDA modeling, and their corresponded perplexities are computed respectively. The optimal value of subject number is selected when the perplexity obtains minimal value. The perplexity distribution diagram is demonstrated in Figure 3.

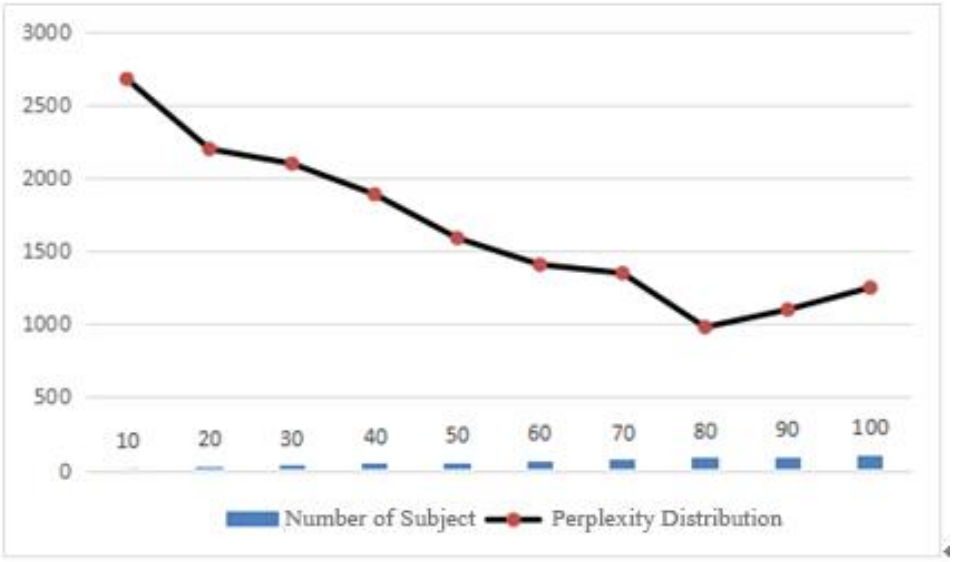

Figure 3. The Perplexity Distribution Diagram

It can be seen from Figure 3 that the perplexity is decreasing gradually when the number of subject is increasing. When the number of subject is equal to 80, the perplexity obtains minimal value, thus the optimal value of subject number is set to k=80. Then 1000 times iteration of Gibbs sampling method is used to obtain the probability distribution matrix DT and TW.

4.3 Experimental Data and Analysis

The experiments use discrete modeling method to analyze the intensity change. LDA model is first use to obtain the subject information from the corpus of science and technology policies, and their released times are extracted so as to divide the policy document into different time range and analysis the change trend with time. Subject 9, 24, and 32 are selected for intensity analysis. The time range is from 2005 to 2015. The change trend of the above three subjects with times is demonstrated in Figure 4.

80 theme in this experiment in modeling to obtain the selected topic 9, 24 and 32 three theme and thematic strength analysis, policy issued time selected in 2005 to 2015 period, time granularity in years as a unit, a formula to calculate the strength of a theme in each time segment derived by utilizing the strength of topics described in section, three topics with time variations in the strength of the trend as shown in Figure 4:

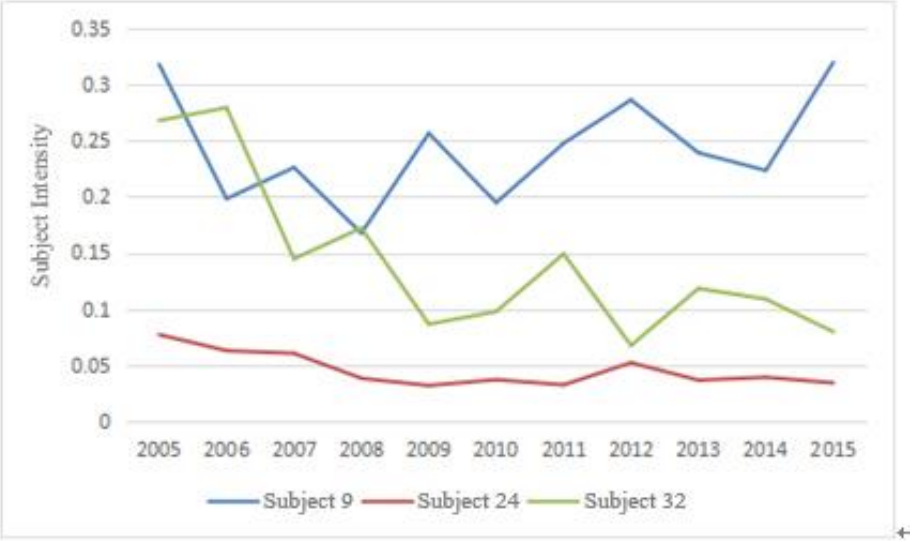

Figure 4. Subject Intensity Analysis under Different Times

It can be seen from Figure 4 that intensity of subject 9 shows an upward trend with time, intensity of subject 32 is gradually decreasing with time, and intensity of subject 24 is with small fluctuation during the ten years. According to the TW matrix of LDA modeling, the top ten key words in the topic 9, 24 and 32 are given in table 3.

Table 3. Some of The Topics and Feature Words

Subject	Feature words
Subject 9	New energy, Biology, Medicine, Environmental Protection, Energy, Electronics, Chemical, Aerospace, Materials, High-tech
Subject 24	Credit, Bank Guarantees, Risk, Business, Financial Institutions, Financing, Loans, Bonds
Subject 32	Software, Computers, Information Systems, Integrated Circuits, Internet, Communications, Information Network, The Software Industry, Electronics, Terminal

From the three subjects given in Figure 4, it can be seen that subject 9 refers to the high-technology industry, subject 24 is the financial support, and subject 32 is the information technology. Combined with area and intensity of subject, it can be clearly seen the change trend of intensity of given subject area in any time range.

In addition, using the implementation scope of science and technology policies, the policy document can be classified into corresponded regions to analysis the intensity change trend under different regions. Subject 9, 24 and 32 are selected as subjects to be analyzed, and implementation regions select "Hebei Province "," Henan Province ", and "Tianjin city" to compute the intensities of the three subjects under different regions shown in Figure 5

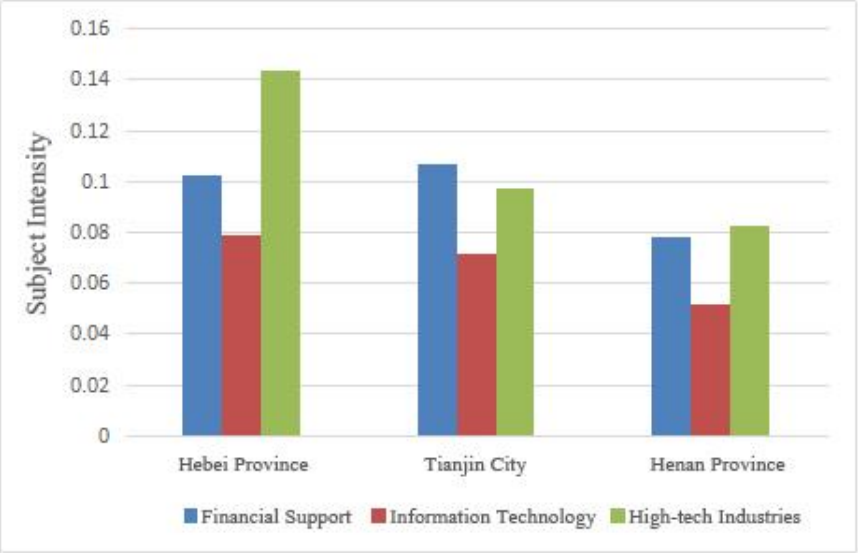

Figure 5 Subject Intensity Analysis under Different Regions

From the above figure, it can be seen that high technology subject obtains the maximum intensity from the science and technology policies of Hebei province, science and technology policy issued by the Tianjin focus on the subject of financial support, and Henan province's science and technology policy is mainly distributed in the financial support and high-tech industries.

5. Conclusion

In this paper, we propose a subject modeling method based on LDA model to extract hidden important subjects for the problem of how to deal with the data set of large scale science and technology policy. According to the particularity of the science and technology policies, we construct subject's library for science and technology policies, and improve the traditional TF-IDE algorithm by focusing on selecting existed words in the library as features in the text weighted feature extraction. In the process of LDA modeling, we choose the optimal number of subjects using standard of perplexity measure for the problem of predefining the number of subjects, propose the concept of Sig(ti), use the released time and the implementation of the scope of information of science and technology policies on the basis of computation of subject intensity, and realize the analysis of intensity change trend of each subject under different conditions of time and region.

Acknowledgement

This study is funded by the National Natural Science Foundation of China (#61373160).

References

[1] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", Expert Systems with Applications, vol. 36, no, 4, **(2009)**, pp. 7764-7772.
[2] A. Zhao, X. Lin and J. Yang, "Graph-based Model for Topic Detection", Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), **(2014)**, pp. 1.
[3] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 993-1022.

[4] M. Shao and L. Qin, "Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence", 2014 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014)). Atlantis Press, **(2014)**.
[5] L. Na, L. M. Xia and L. Ying, "Mixture of topic model for multi-document summarization", Control and Decision Conference (2014 CCDC), The 26th Chinese. IEEE, **(2014)**, pp. 5168-5172.
[6] J. F. Wang, K. X. Shen, A. F. Xu and Y. L. Lan, "An Improved Lda Model in Micro-Blog Tags Extracting Based on Multi-Tags", The Open Cybernetics & Systemics Journal, vol. 8, **(2014)**, pp. 1266-1270.
[7] H. C. Wu, R. W. P. Luk and K. F. Wong, "Interpreting TF-IDF term weights as making relevance decisions", Acm Transactions on Information Systems, vol. 26, no. 3, **(2008)**, pp. 55-59.
[8] A. Asuncion, M. Welling and P. Smyth, "On Smoothing and Inference for Topic Models", Twenty-fifth Conference on Uncertainty in Artificial Intelligence, **(2009)**, pp. 27-34.

Authors

Jianmin Wang
（1975－）,male, associate professor, research direction: software engineering, document clustering. Email: blacklion75@126.com.

Tongrang Fan
（1965－）, female, professor, research direction: network technology and security, information processing, network and education. Email: fantr2009@126.com.

