Keywords Extraction Method for Technological Demands of Small and Medium-Sized Enterprises Based on LDA

1st Xingbing Liu
College of Computer and Information Engineering
Henan Normal University
Big Data Engineering Laboratory for Teaching Resources and Assessment of Education Quality ,Henan Province
Xinxiang, China
liuxingbing@htu.edu.cn

2nd Zhen Zhang
College of Computer and Information Engineering
Henan Normal University
Xinxiang, China
412015806@qq.com

3rd Baoxia Li
College of Computer and Information Engineering
Henan Normal University
Xinxiang, China
2939262407@qq.com

4rd Fei Zhang
College of Computer and Information Engineering
Henan Normal University
Xinxiang, China
zhangfei@htu.edu.cn

***Abstract*—Keywords extraction technology is used in the technological demands of SMEs. It can match the demands to the scientific research team quickly and accurately. This is an effective method to promote industry-university-research linkage. But how to extract effective information from a wide range of technological demands information is a challenging task. For this issue, the paper is based on the text of the technological demands of SMEs, and proposes a keywords extraction method based on LDA. Firstly, natural language processing technology is used to preprocess text. Then, multi-feature weighting and Latent Dirichlet Allocation (LDA) theme model are fused to extract keywords. Thirdly, the method chooses the best keywords through post-processing. Finally, the method is verified by experiments, and the feasibility of the method is demonstrated by actual cases. The method was tested on Chinese ScientistIn dataset. The experimental results show that the F-value is higher than that of Term Frequency - Inverse Document Frequency (TF - IDF) and LDA. It also shows that the method can improve the intelligence of the scientific and technological collaboration platform. The algorithm works best when the *F*-value is 0.79 at *K*=3.**

***Index Terms*—keywords extraction, LDA theme model, text processing**

I. Introduction

The status of SMEs in the economic development of society is very important. The pressure of technological upgrading of SMEs and the pressure of its comprehensive development have emerged with the background of technological innovation [1]. The weakness of SMEs lies in the fact that the reserves of their high-tech talents are insufficient, and on the other hand, the development of science and technological teams is not advanced enough, so the competition for talents is at a disadvantage. When a company encounters a problem, it is usually through an entrepreneur using his or her own social resources to find an expert or research team to solve the problem. Obviously, the level of resources owned by entrepreneurs determines the effectiveness of the solution to the problem, which is very unfavorable for SMEs. Therefore, it is possible to combine the demand information of SMEs technology with the online platform of technology collaboration. The method is a well strategy to solve the intelligence of enterprise demand information. And demand information can be recommended to universities, research institutes or research teams accurately.

A collection of keywords is usually a few words or phrases that it can concise the subject of the text highly. Some people can manage, retrieve and read documents with keywords information efficiently [2]. Usually in the process of project development of the technological platform, the keywords of the demands are filled or marked by the user. Obviously, there are subjective opinions in this situation. In addition, the number of keywords marked by each user is different, and the keywords of the mark may be missing or one-sided. What's more, some platforms do not have a keywords column. Manual tagging is used when dealing with keywords extraction. Because of the large amount of corpus information, the work of marking becomes difficult. Therefore, it brings difficulties and challenges to scholars. Automatic extraction of keywords means that the extraction of keywords is done uniformly using a computer. The method can grasp the theme

This work is supported by the National Natural Science Foundation of China under Grant Nos. U1804164.

or key points of the document quickly, and it is also the content of many scholars [3].

The paper combines the topic model with the information text of the technological demands of SMEs. The method LDA theme model is based on the multi-feature weighted sorting of keywords, and the keywords extraction of information for SMEs' technological demands is realized. In the end, the experimental results are superior to the traditional topic model algorithms. The rest of the paper is organized as follows. Section 2 describes the closely related works. Section 3 details the architecture of our keywords extraction system (base on improve LDA). Section 4 evaluates our models with dedicated experiments and Section 5 concludes the paper.

II. RELATED WORK

Regarding the demands of enterprises, Liu Y F et al. [4] proposed a method for enterprises to achieve the goals of maximizing customer demand and minimizing costs. This method uses the NSGA-II genetic algorithm. And it is aimed at the problem of order processing of equipment manufacturing enterprises. Li Ying [5] used the topic model to build a vector space model of the technical demands of the enterprise, and then matched it with experts. However, the focus of this method is recommendation and analysis on the expert side. Yu Juntao [6] proposed the method of using content vectors and semantic vectors to represent experts' scientific research literatures and process related data sets. The method was applied to the research of industry-university-recommended recommendation algorithms and recommendation systems. It is a pity that the method focuses on processing scientific research documents. Kang J et al. [7] combined LDA and clustering algorithms to determine the best matching team after classifying enterprise technology categories. The method can realize the choice of the partner of industry, university and research, but it ignored the analysis on the enterprise side. To sum up, researchers have rarely studied the characteristics of enterprise demand texts, and most of them are unstructured Chinese texts, so the accuracy of demand feature extraction is relatively low.

In terms of topic models, the Term Frequency - Inverse Document Frequency model is one of the earliest text probability models. In the corpus, the word frequency (TF) is the quotient of the number of occurrences of a word and the total number of words in the document. The inverse document frequency (IDF) represents the logarithm of the total document in the corpus and the quotient of the number of documents containing the word. The result of multiplying the word frequency and the inverse document is the TF-IDF value [8]. The disadvantage of this model is that it only judges whether the word is a key word by the frequency of its occurrence. Latent Semantic Indexing (LSI) model decomposes singular value (SVD) based on TF-IDF. It maps the word vectors in the corpus to a lower dimension space, and also takes into account the meaning problem [9]. However, the disadvantage is that the time complexity of SVD for high-dimensional matrices is quite different, and the choice of topic number has a great impact on the results. The Probability Latent Semantic Indexing (PLSI) model is more optimized than LSI. PLSI adds an implicit topic layer on the basis of LSI. The implicit topic is generated by words according to a certain probability, and the semantics of variables are related to each other [10]. Although it solves the complex characteristics of LSI in high-dimensional processing, it is prone to over-fitting problems with the increase of the number of texts. Blei (2003) proposed the LDA model based on PLSI [11]. Dirichletting the parameter variables existing on the PLSI model better overcomes the problem of over-fitting and linear increase. In contrast, the LDA model is a complete probability generation model, which attracts more attention and wide application of researchers. The paper aims at unstructured text information about the technological demands of enterprises, and mines key information characteristics based on LDA.

III. RESEARCH IDEAS AND MODELS

A. Research ideas

The process of keywords extraction is usually divided into four steps: text preprocessing, candidate set, keywords classification and extraction, and post-processing. Text preprocessing can be divided into information sampling, partial sample conversion or translation, segmentation of words, and data cleaning. The selection of candidate sets is to identify the important sentences or words in the corpus. The classification or extraction of keywords is to sort the weight or probability of keywords by the algorithm of keywords extraction, and then extract or classify the top-N ranking keywords. Post-processing is a method of merging adjacent keywords or phrases to ensure that the extracted keywords meet the best requirements of the researcher.

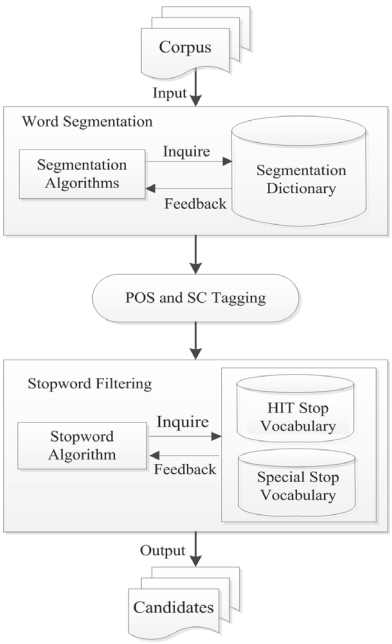

Fig. 1. Preprocessing diagram.

In order to extract the keywords of the technological demands of the enterprise effectively , the paper retrieves and collects data from the database according to the project classification. Then the initial data is pre-processed, which includes the process of cleaning the corpus, word segmentation, part-of-speech annotation and stop words filtering. Then the preprocessing of the initial data includes corpus cleaning, word segmentation processing, part-of-speech annotation and stop words filtering. The author cleans and deletes bad data including extreme speech, illegal language and test data. In order to improve the effect of word segmentation, a special dictionary is set up as a user dictionary to be added to the original word segmentation tool, so that the segmentation result of the method is more accurate. The paper uses Harbin Institute of Technology and proprietary stop words lists for stop words filtering. Among them, the proprietary stop words list is a set of stop words in a specific field, which are established after statistics and analysis. The paper uses the LDA topic model to extract keywords, combines feature weighting and the LDA topic model algorithm to sort the keywords, and then performs post-processing, combining adjacent words into a single readable phrase to obtain the best keywords, and recommend to the corresponding scientific research team accurately. Finally, the *F*-value of the evaluation comprehensive index is used to analyze the accuracy of different algorithms, and the accuracy of the algorithm in the paper is verified.

B. Models

The LDA model is a topic model for potential topics that are not clearly shown in the generated document. In the LDA topic model, select a topic with a certain probability in the document, and select a word with a certain probability from the topic [12]. The three-layer Bayesian structure representation of the model is shown in Figure 2. The *K*, *M*, and *N* are the number of loops, which represent the number of topics, documents, and words in the document respectively. w represents a word, θ is the topic probability of the document, φ is the word probability of the topic, z is the topic label assigned to the word w[13]. Zhou Zhihua [14] stated in Machine Learning that the topic structure of the model can be solved by formula (1). Because the denominator in (1) is difficult to obtain, the learning methods of θ and φ parameters usually use Gibbs sampling or variational method to infer approximately. Scholars usually prefer Gibbs, which is easy to implement and time-saving in large-scale text set [15].

$$p(z, \varphi, \theta | w, \alpha, \beta) = \frac{p(w, z, \beta, \theta | \alpha, \beta)}{p(w | \alpha, \beta)} \tag{1}$$

C. parameters selection

In the training of LDA model, the empirical value of super parameters α and β is α = *K*/50, β = 0.01 [16]. For the determination of the number of model topics *K*, one case is through the training results of *F*-value or confusion degree, and another case is to manually specify when part of the text

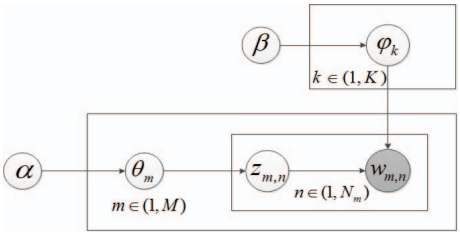

Fig. 2. LDA Topic Model Diagram.

is small. The calculation method of *F*-value training method is shown in formula (2), in which *P* stands for the accuracy rate and *R* stands for the recall rate. The corresponding content will be introduced in detail in the experiment part of the following section 4. The calculation method of perplexity evaluation method is shown in formula (3). The perplexity decreases with the increase of *K*, and *K* is the best when the trend of perplexity is no longer obvious or stable.

$$F = \frac{2PR}{P + R} \tag{2}$$

$$perplexity(d) = \exp(-\frac{\sum \log p(w)}{\sum_{d=1}^{M} N_d}) \tag{3}$$

IV. DESCRIPTION OF THE ALGORITHM

A. Corpus composition

Part of the data of the technological demands information of SMEs is the background data of the author's horizontal project platform. It contains information such as demand title, detailed demand, limited time, basic budget situation and so on. The title of the demand is an overview of the most streamlined situation. The detailed demand contains a lot of detailed information, and it is also an important corpus for obtaining keywords. The time limit is the estimated time cost of completing the project. The basic budget situation is an approximate budget for the funding of the task. Another part of the data is web crawler data from Scientists Online (www.scientistin.com). From the content point of view, most of the data does not have a specific format, the necessary title part and detailed demand is not indispensable. Therefore, from the text of the enterprise demand information, the title and detailed demand is an important part of the keywords extraction of the enterprises demand. In order to facilitate the grasp of the demand topic information and reduce the time complexity of the model, the title and detailed demand is combined into one document when processing the data.

B. Description of multi-feature LDA model

The process of LDA model training is to make each word in the text go through the process of "selecting a certain topic with an accurate probability, and then selecting a certain word from this topic with a certain probability" [17]. According to formula (1), the traversal rule is that the product of the probability of a word under the same topic and the probability

of belonging to that topic under the same document is equal to the frequency of the word in the document. The LDA model trained on a large amount of data can obtain the topic distribution of the document and the word distribution of the topic.

The LDA topic model contains three layers of Bayesian structure: words, topics, and documents. A document can be regarded as a collection with many topic distributions, and a set of many words constitutes a certain topic. However, there are some general characteristics of words and sentences in the text of the corpus, for example, sentences are a collection of words, words with different parts of speech tags have different contributions to sentences, and words in different sentence components have different contributions to sentences, etc. [18]. Therefore, the parts of speech and components of the extracted words affect the semantic expression of the document.

The paper considers the degree of influence of part-of-speech and component features on sentences, and then weights the features before ranking the keywords. The initial keywords extracted by the LDA model are represented by w, and the distribution probability of w is $p(w|\alpha, \beta)$. For the convenience of the reader, the following are expressed in terms of $p(w)$, where $p(w_{m,n})$ is the probability of the *n*-th keyword of the *m*-th document. $pos(w)$ represents the part-of-speech weight of w. Similarly, $pos(w_{m,n})$ is the part-of-speech weight of the *n*-th keyword of the *m*-th document. The analysis result of the keyword part-of-speech tag weight in [18] is shown in the formula (4). $sc(w)$ represents the sentence component weight of w. Similarly, $sc(w_{m,n})$ represents the sentence component of the *n*-th keyword of the *m*-th document. The analysis result of the keyword sentence component weight in [19] is shown in the formula (5). The total probability distribution of the keywords weighted by the fusion feature is represented by t, $t_{m,n}$ is the total weight of the *n*-th keyword of the *m*-th document, and the calculation formula is $t_{m,n} = p(w_{m,n}) \times (pos(w_{m,n}) + sc(w_{m,n}))$.

$$pos(w_{m,n}) = \begin{cases} 0.8 & if\ w_{m,n}\ is\ noun \\ 0.5 & if\ w_{m,n}\ is\ verb \\ 0.4 & if\ w_{m,n}\ is\ adj \\ 0 & if\ w_{m,n}\ is\ others \end{cases} \tag{4}$$

$$sc(w_{m,n}) = \begin{cases} 0.5 & if\ w_{m,n}\ is\ subject \\ 0.2 & if\ w_{m,n}\ is\ predicate \\ 0.3 & if\ w_{m,n}\ is\ object \\ 0 & if\ w_{m,n}\ is\ others \end{cases} \tag{5}$$

V. EXPERIMENT

A. Experimental environment

The experiment was performed on a computer with 8G memory and a Windows 10 system. The experimental training and testing uses Python 3.7 version, and calls the lda package in the gensim library to implement the LDA algorithm. The crawler uses the Python third-party module requests to grab data, and through page rotation to obtain a list of items on each page of the website. And then use regular expressions to crawl the homepage URL corresponding to each item list. Further retrieve and grab the project title and project content description on each project homepage, and save these data in an Excel spreadsheet. In the experiment, the jieba library is used as the word segmentation tool, and the jieba.posseg.dt default part-of-speech tagging tokenizer in the jieba library is used to label the part of speech. After data cleaning, the experimental extraction project requires a total of 300 corpora in six categories, some of which contain manually tagged keywords. In the experiment, 200 pieces of data were used as the training set, and the number of topics in the training model was *K*. The other 100 data sets are used as the test set as the basis for evaluating the algorithm in the paper. For the data set, each document was extracted with keywords by three humans. In addition, based on the same test set, the three algorithms of the paper, TF-IDF model and traditional LDA model are used for cross comparison.

TABLE I
THE FLOW CHART OF THE ALGORITHM.

Step	Algorithmic flow
Step 1	Getting text
Step 2	Text Preprocessing to Get Candidate Sets
Step 3	Input α=50/*K*, β=0.01
Step 4	for m in Top
Step 5	for Top in z
Step 6	Repeat steps 4 and 5 until each word is traversed
Step 7	Probability Distribution of Document-Theme-Word
Step 8	Output $Top - N$ Ranking of Probabilities after Weighting
Step 9	END

B. Evaluation criteria

The extraction of keywords for enterprises' demand is to select a set of words that can summarize the key information of the enterprises' demand document. In terms of the definition and intrinsic meaning of keywords, the best evaluation criterion for the effectiveness of extracting keywords from text is whether the keywords themselves conform to the actual theme and semantics of the document. In terms of scientific research and academic perspectives of keywords, the strategies for evaluating the extracted keywords include the conditions of judging whether the structure is stable and the semantics are complete, and further better information mining of text [20]. Inspired by the LDA topic model, the paper uses multi-feature weighting to reorder keywords and extract them, and proposes a method for keywords extraction based on the technological demands of SMEs based on LDA.

At present, most researchers use the accuracy rate *P* (Precision), the recall rate *R* (Recall), and the combined value of the *F*-value (quantitative evaluation) to evaluate the topic model. The calculation formulas for *P R* and *F* are (6) (7) and (2). *TP* indicates that the prediction is positive and the actual is positive. *FP* indicates that the prediction is positive and actually negative. *FN* indicates that the prediction is negative and actually positive. The accuracy rate *P* represents the proportion of the extracted correct keywords to the number of

TABLE II
EXPERIMENTAL RESULTS OF THE THREE ALGORITHMS.

Direction of Demand	TF-IDF			LDA			THE Algorithm		
	P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Environmental Protection	0.42	0.42	0.42	0.40	0.60	0.48	0.49	0.74	**0.59**
Biomedicine	0.60	0.60	0.60	0.53	0.79	0.63	0.57	0.86	**0.69**
Machine Made	0.54	0.54	0.54	0.41	0.61	0.49	0.52	0.78	**0.62**
Chemical Industry	0.53	0.53	0.53	0.44	0.65	0.52	0.49	0.74	**0.59**
New Energy	0.49	0.49	0.49	0.40	0.60	0.48	0.50	0.75	**0.60**
Material Science	0.63	0.63	0.63	0.54	0.82	0.65	0.57	0.85	**0.68**

extracted keywords. The recall rate R represents the proportion of the correctly extracted keywords in the manually labeled keywords in the sample. The F-value is a comprehensive evaluation of both.

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

C. Experimental results and analysis

The data set required by the experimental enterprises has six directions, energy conservation, environmental protection, biomedicine, and machinery manufacturing. The parameter K affects the experimental accuracy of the LDA model and the algorithm in the paper. The accuracy of the TF-IDF algorithm is affected by the number of keywords. Therefore, the experiment uses the principle of controlled variables to conduct experiments on related data. Table II is completed under the condition of K = 3 and two keywords for each topic (n = 6), and the corresponding number of TF-IDF algorithm keywords is n = 6. It is guaranteed that the number of keywords extracted by each model in the experiment is 6. In order to facilitate comparison and calculation, the number of keywords manually labeled in each demand document is 5. The values of the hyperparameters α = K/50, β = 0.01. The experiments were performed on the basis of the above data sets and parameters.

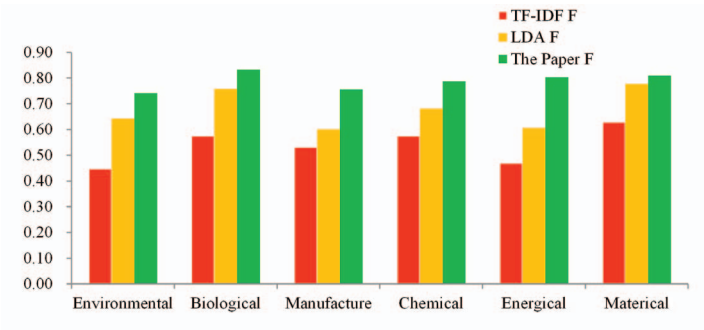

Fig. 3. Comparison of F-values of three algorithms.

The results in Table II and Figure 3 were all performed at n = 6. According to the experimental results, it can be seen that the F-values of the algorithm in the six research directions are 0.59, 0.69, 0.62, 0.59, 0.60, 0.68 in this order. It can be seen from Figure 3 that the F-value of each research direction of the algorithm in the paper is numerically higher than the other two algorithms. Moreover, Table II also shows that the P and R values of the algorithm in the paper are also higher than the other two algorithms. So it also directly shows that the algorithm in the paper is superior to the commonly used TF-IDF and traditional LDA algorithms.

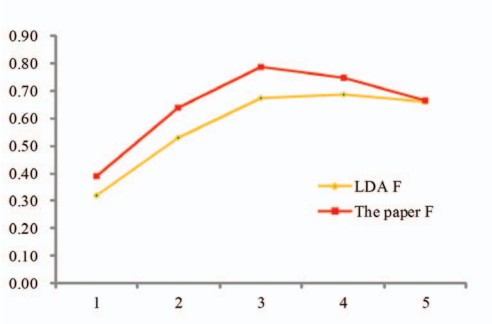

Fig. 4. Comparison of F values at different K.

In addition, the number of topics K that has a significant effect on the experimental results determines the number of keywords extraction. In addition, the setting of different distances has an impact on the accuracy of keywords extraction. The experiment set K to be an integer between 1 and 5, and the number of keywords under each topic is 2 to ensure the quantity and quality of keywords extraction, and the best K value is obtained through training. Figure 4 shows the F-value of the algorithm and the traditional LDA model as K changes. The F-value of the algorithm in the paper is higher than the traditional LDA model between 1 and 5. However, as the number of topics K increases, the F-value of the two algorithms gradually approaches, and the F-value tends to rise first and then decrease. The reason is that with the increase of K, the larger the number of keywords extracted in the model, the FP in formula (5) gradually increases, so the accuracy rate P is gradually decreasing, and the F value is also continuously decreasing. The algorithm in the paper works best when $K = 3$.

TABLE III
EXTRACTION RESULTS OF THREE ALGORITHMS.

Methods	Results
Annotate results manually	Coal,Vermiculite,Recognition system, Dry separation equipment,Dry separation, equipment,Recognition rate
TF-IDF	Coal,Vermiculite,Dry election,Wash, Downhole,intelligent
Traditional LDA algorithm	Equipment,Smart, Research and development,Dry election, Coal,Vermiculite
The algorithm	Coal,Vermiculite, Dry separation equipment, Intelligent identification system, Recognition rate,Wash

D. Case scenarios and applications

In order to observe the feasibility of the method in the paper on keywords processing, the author takes the demand information of an enterprise as shown in Figure 5, and then organizes the demand keywords extracted by the three methods into Table III. From Table III, the TF-IDF model algorithm is based on word frequency, and the accuracy of the output results is relatively poor. The LDA model is closer to the original annotation results semantically. The effect of the fusion of multi-feature weighting is obvious, but some stop words and synonyms affect the extraction effect, which can be improved in future research.

煤炭、矸石的智能识别系统需解决的问题：鉴于环保、节能的要求作为主要能源及重要工业原料的煤炭，为提高其煤质，对选煤工业的要求越来越高。市场现在广泛采用专职选煤厂重介洗选的工艺，这样既造成水资源的浪费，同时地面矸石的堆放也造成环境的污染和土地的浪费。因此我们在传统洗选设备的基础上致力于井下煤炭干选设备的研发。技术难题：井下干选设备中，煤炭、矸石的智能识别系统研发，使智能识别率在90%以上。

Fig. 5. Example of demand information.

CONCLUSION

In the traditional offline processing method, if SMEs encounter difficulties in technological demands, they seek help through the personal strength of entrepreneurs. The problems of the enterprise could not be solved in time, and no relevant technological innovation information was shared. The demand information of enterprise is recommended to universities, research institutes or scientific research teams accurately through technology platforms. The approach saves time and effort in practical applications, and is conducive to the scientific and technological progress and development of enterprises, as well as the realization of industry-university-research linkage. The method proposed by the paper combined with the platform has improved the intelligence of the platform significantly.

The paper starts with the technological demands text of the enterprise, and proposes an LDA topic model algorithm that integrates multi-feature weighting. Compared with the traditional algorithm, the algorithm also has a significant improvement in accuracy. In future research, we will improve from the following research directions. We will better improve and perfect the LDA theme model. We will further improve the operating mechanism of the technology collaboration platform for SMEs. We will use the model in other applications Field promotion, such as information model of scientific research team, expert recommendation system, etc.

REFERENCES

[1] Y.R Zeng, L Wang, X.H Xu. “An integrated model to select an ERP system for Chinese small- and medium-sized enterprise under uncertainty.” Technological and Economic Development of Economy 23.1 (2017): 38-58.

[2] Q Qiu, Z Xie, L Wu, et al. “Geoscience keyphrase extraction algorithm using enhanced word embedding”. Expert Systems with Applications, 125 (2019): 157-169.

[3] J.W Fang, H.R Cui, G.X He, et al. “Academic Text Keyword Extraction Based on Prior Knowledge TextRank”. Information Science 3 (2019): 13.

[4] Y.F Liu, Q.S Zhang. “Solving multi-objective planning model for equipment manufacturing enterprises with dual uncertain demands using NSGA-II algorithm”. Advances in Production Engineering and Management 13.2 (2018): 193-205.

[5] Y Li. Research on Expert Recommendation Algorithm for Enterprise Demand. Diss. Beijing: Beijing Jiaotong University. 2018.(Chinese)

[6] J.T Yu. Research on recommendation algorithm and recommendation system for industry-university-research. Diss. Nanjing: Southeast University, 2017.(Chinese)

[7] J Kang, J Lee, D Jang, et al. “A Methodology of Partner Selection for Sustainable Industry-University Cooperation Based on LDA Topic Model”. Sustainability, 11.12 (2019): 3478.

[8] K Chen, Z Zhang, J Long, et al. “Turning from TF-IDF to TF-IGM for term weighting in text classification”. Expert Systems with Applications An International Journal, 66.C (2016): 245-260.

[9] S Deerwester, S.T Dumais,G.W Furnas, et al. “Indexing by latent semantic analysis”. Journal of the American society for information science, 41.6 (1990): 391-407.

[10] T Hofmann.“Unsupervised Learning by Probabilistic Latent Semantic Analysis”.Machine Learning, 42.1-2 (2001): 177-196.

[11] D.M Blei, A.Y Ng, M.I Jordan. “Latent Dirichlet Allocation”. Journal of Machine Learning esearch, 3.Jan (2003): 993-1022.

[12] Y.L Zhang, F.E Christoph. “Tracking Events in Twitter by Combining an LDA-Based Approach and a DensityCContour Clustering Approach”. International Journal of Semantic Computing, 13.1 (2019): 87-110.

[13] X GUI, J ZHANG, X ZHANG, et al. “Survey on Temporal Topic Model Methods and Application”. Computer Science, 2.(2017): 6.

[14] Z.H Zhou. “Machine Learning”[M]. Beijing: Tsinghua University Press, (2016): 337-340.(Chinese)

[15] Z Qiu, B Wu, B Wang, et al. “Gibbs Collapsed Sampling for Latent Dirichlet Allocation on Spark”. Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, (2014): 17-28.

[16] H Jelodar, Y Wang, C Yuan, et al. “Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey”. Multimedia Tools and Applications, 78.11 (2019): 15169-15211.

[17] K Bastani, H Namavari, J Shaffer. “Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints”. Expert Systems with Applications, 127 (2019): 256-271.

[18] Z.H Wang, Y Guo. “Sentence-Ranking-Enhanced Keywords Extraction from Chinese Patents”. J. Inf. Sci. Eng., 35.3 (2019): 651-674.

[19] Z.X Zhang. Modular Chinese sentence similarity calculation based on HowNet. Diss. Ma Anshan: Anhui University of Technology, 2010.(Chinese)

[20] Siddiqi, Sifatullah , A Sharan. “Keyword and Keyphrase Extraction Techniques: A Literature Review.” International Journal of Computer Applications 109.2 (2015): 18-23.