Modern Applied Science; Vol. 13, No. 1; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

# Study on the Topic Mining and Dynamic Visualization in View of LDA Model

Ting Xie[1], Ping Qin[1,2] & Libo Zhu[1]

[1] Economics and Management, Nanjing University of Aeronautics & Astronautics, 210000, Nanjing, China

[2] Evaluation Center for Think Tank of Industry and Information Technology, China

Correspondence: Jiangjun Rd. Campus: 29 Jiangjun Ave., Nanjing, 211100, Nanjing. Tel: 1-515-067-6279. E-mail: chilli6279@163.com

Received: October 19, 2018 Accepted: November 1, 2018 Online Published: December 31, 2018

doi:10.5539/mas.v13n1p204 URL: https://doi.org/10.5539/ mas.v13n1p204

*The research is financed by the basic scientific research business fee of the central university, This research is one of the research results of the project "Research on Performance Evaluation and Application of University Think Tanks Based on Deep Learning" (No. NK2018009).*

**Abstract**

Text topic mining and visualization are the basis for clustering the topics， distinguishing front topics and hot topics. This paper constructs the LDA topic model based on Python language and researches topic mining，clustering and dynamic visualization，taking the metrology of Library and information science in 2017 as an example. In this model，parameter and parameter are estimated by Gibbs sampling，and the best topic number was determined by coherence scores. The topic mining based on the LDA model can well simulate the semantic information of the large corpus，and make the corpus not limited to the key words. The bubble bar graph of the topic-words can present the many-to-many dynamic relationships between the topic and words.

**Keywords:** dynamic visualization, LDA model, metrology, Python, topics mining

## 1. Introduction

Text content mining and semantic modeling are the research hotspots and core content in the field of information recommendation and data mining [1]. Mining text themes is a key and basic work in the field of text mining. Through text mining, the text is reduced from word space to topic space. Under the current network environment, information content has dynamic interaction and evolution with time. Therefore, it is of great significance to innovate information mining methods and visualize the dynamics of information. This article will conduct LDA modeling through Python, implement topic mining in depth at the abstract level, and dynamically visualize it through Python's pyLDAvis package and jupyter.

## 2. Related Research Status

In 2013, Shi Qiankun studied text topic mining and text static visualization analysis based on LDA model[2]. In 2014, Hu Jiming and other LDA models for dynamic text content mining explored the evolution of content topics over time from the perspective of topic similarity and intensity[1];Jing Chen et al. The 1990-2014 literature included in the WOS database The data is the research object, using Cite Space for data visualization analysis, showing the power distribution and journal distribution of topic mining research, and analyzing the key academic literature supporting the topic mining[3]. In 2015,Liu Jinhua used the word frequency statistics and keyword clustering method based on co-occurrence relationship to construct the map, and visualized the clustering results through the visualization tool VOSviewer[4]; Wang Peng et al. based on the LDA model, estimated the text theme through Gibbs algorithm. Probability distribution, using JS (Jensen-Shannon) distance as the similarity measure of text, using hierarchical clustering method for clustering. The validity of the method is proved by clustering purity and Fscore value[5]. In 2016, Wang Yufen and others used a method based on the quantitative analysis of the recall rate, the precision rate, the F value and the information entropy, and the qualitative analysis of the breadth of the subject matter extraction and the subjective granularity. The LDA theme extraction effect is better than the keyword and abstract as the corpus LDA theme extraction effect[6]; Yang Chao et al. based on the patent-structured SAO (subject-action-object (SAO) structure LDA model to achieve the patent subject structure Identification and

analysis[7]. The 2017 LDA theme model is widely used in topic recognition, topic clustering, and topic mining, trend analysis, and correlation analysis.

In summary, the commonly used topic mining and visualization research tools are CiteSpace and VOSviewer. The related papers based on LDA model research mainly focus on topic mining and topic clustering of text data in a certain field. The visualization is mainly document- Static rendering of two matrices of topics, topics, and vocabularies, dynamic visualization analysis has not been covered. This article will use the LDA model to model the topic mining and dynamic visualization of the 2017 library information field through Python modeling.

## 3. Title Page Subject Recognition Based on LDA Model

### *3.1 Basic Principles of the LDA Model*

In 2003, Blei et al. proposed the latent Dirichlet Allocation model, which is a three-layer Bayesian probability generation model, which uses iterative estimation to calculate the subject vocabulary of the document[8]. The main idea is to assume that each document is a mixture of multiple topics, and each topic is a probability distribution over multiple vocabularies. There are D-documents and W-vocabulary in the corpus. Assuming that the documents have a K-topic, the process of generating the LDA model theme is:

(1) For each document dϵD, according to the Dirichlet distribution $\theta_d \sim Dir(\alpha)$, the subject distribution parameters of the document d are obtained;

(2) For each topic zϵK, according to the Dirichlet distribution $\varphi_z \sim Dir(\beta)$,the multi-distribution of the vocabulary on the subject z is obtained;

(3) For the i-th vocabulary in document D, the subject is obtained according to the polynomial distribution $Z_{d,i} \sim Mult(\theta_d)$.

In the LDA model, parameter settings, parameters, and most empirical studies are based on the rule of thumb, ie, setting$\alpha=50/K$, $\beta=0.01$; the number of topics K is constrained by the subject consistency score, that is, when the consistency score is the highest, K takes the most Excellent value; two Dirichlet distributions $\theta_d$、$\varphi_z$, which cannot be directly obtained, are estimated by Gibbs sampling algorithm in actual research.

### *3.2 Principle of Topic Recognition Based on LDA Model*

The topic contribution rate is the weight of the topic in the document collection and is a quantitative indicator of the research topic. The topic contribution rate can be expressed by the ratio of the sum of the weights of the research topic in the document collection to the volume of the document.

For each topic, the formula for calculating lexical saliency based on the word frequency and the conditional probability distribution of the subject in the vocabulary is as follows:

$$Saliency(w)=frequency(w)\left[sum_t P(t|w) * log\left(P(t|w)\right)/p(t)\right] \quad (1)$$

The formula for calculating the relevance of the vocabulary for each topic distribution is as follows:

$$Relevance(w|t)=\lambda * P(w|t) + (1-\lambda) * P(w|t)/P(w) \quad (2)$$

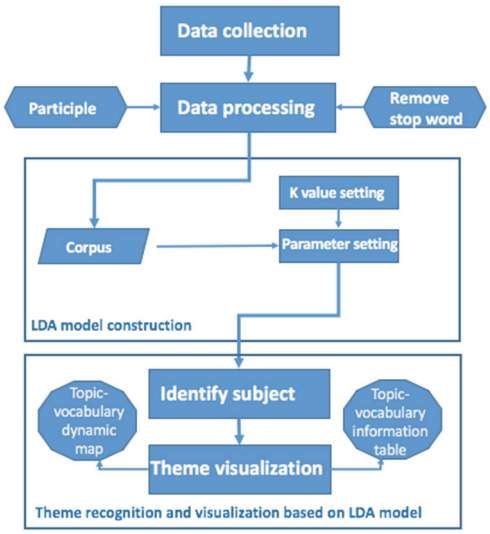


Figure 1. topic identification and visualization flow chart

In equations (1) and (2), t represents the subject, w represents the vocabulary, P is the probability, Saliency is significant, frequency is frequency, and Relevancy is correlation. According to formulas (1) and (2), the pyLDAvis function can be called in Python language to get a bubble-bar graph of the dynamic visualization of the subject-vocabulary. One of the bubbles represents a topic and the bar chart is the vocabulary distribution under the topic. Different subject bubbles correspond to different vocabulary bar charts, and different words correspond to different subject bubbles.

*3.3 Based on LDA Model Theme Recognition Process*

LDA-based research topic recognition mainly includes: data acquisition and processing, extraction based on LDA model theme and identification of cutting-edge topics. The identification process is shown in Figure 1.

Data acquisition and processing stages. First, the acquired data is divided into words, deactivated words, and other data processing operations. The word segmentation uses Python's jieba data package; the stop word list uses the Harbin Institute of Technology stop vocabulary, Sichuan University Machine Intelligence Laboratory stop word list and Baidu stop vocabulary to de-emphasize, the stop word will also use Python get on.

Construction and recognition based on the theme of the LDA model. LDA modeling is performed by Python language. The Gibbs sampling algorithm is used to estimate the topic distribution parameters of the document d and the multi-distribution of the vocabulary on the topic z. The consistency number is used to select the optimal number of topics, and the topic number consistency score curve is drawn. Find the optimal number of topics and perform topic recognition in the case of optimal values.

Visualization based on the theme of the LDA model. The pyLDAvis package in Python is used in conjunction with the jupyter notebook to implement a dynamic visualization of the topic-vocabulary and to plot a breakdown of the topic-related information for each article.

## 4. Empirical research

*4.1 Data Source*

In the Chinese Knowledge Network database search, the chapter titled “Metrics” or “Metrics” was published in 2017. The subject was a library of information and digital libraries. A total of 857 results were retrieved. The title, author, keyword, abstract and time fields of the results were extracted as the data source for the study.

*4.2 Data Processing*

The word segmentation uses Chinese word segmentation component jieba, which supports three word segmentation modes: precise mode, trying to cut the sentence most accurately, suitable for text analysis; full mode, scanning all the words in the sentence that can be worded, very fast , but can not solve the ambiguity; the search engine model, based on the precise mode, split the long words again, improve the recall rate, suitable for search engine segmentation. The corpus based on this paper is in the form of text, so the text in each time slice is segmented by the precise mode. The specific code is shown in Figure 2, and the result is shown in Figure 3.

```
# -*- coding: cp936 -*-
import jieba
f=open(r"\\Mac\Home\Desktop\Keywords-Abstract-Time\keywords-abstract-2017.txt","r")
s=f.read()
seg_list = jieba.cut(s,cut_all=False)
print ('/ '.join(seg_list))
```

Figure 2. Python language word segmentation code

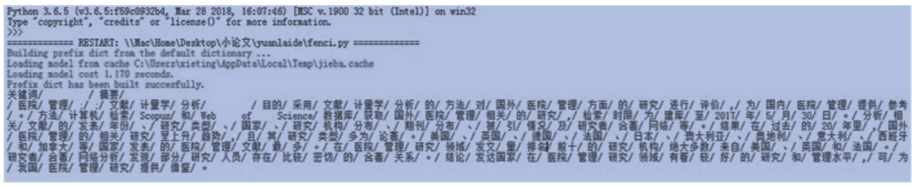


Figure 3. word segmentation results

To stop words: remove punctuation, function words, quantifiers, conjunctions and other words that have no practical meaning. This article uses Baidu stop vocabulary, Harbin Institute of Technology stop word list and Sichuan University Machine Intelligence Laboratory stop word list and Baidu's preposition table, conjunction table,

quantifier. Pre-run the results, the meaningless words involved are manually added to the deactivation table, and later added; distribution, China, domestic, utilization, this article, reveal, etc., to stop the word Python code as shown in Figure 4. Show.

```
def stopwordslist(filepath):
    stopwords=[]
    for line in open(filepath,'r').readlines():
        stopwords+=line.split()

    #stopwords = [line.split() for line in open(filepath,'r').readlines()]
    return stopwords

stopwords0= stopwordslist(r'\\Mac\Home\Desktop\停用词处理\百度.txt')
stopwords1= stopwordslist(r'\\Mac\Home\Desktop\停用词处理\哈工大停用词表.txt')
stopwords2= stopwordslist(r'\\Mac\Home\Desktop\停用词处理\四川大学机器智能实验室停用词库.txt')
stopwords3= stopwordslist(r'\\Mac\Home\Desktop\停用词处理\介词.txt')
stopwords4= stopwordslist(r'\\Mac\Home\Desktop\停用词处理\连词.txt')
stopwords5= stopwordslist(r'\\Mac\Home\Desktop\停用词处理\量词.txt')
stopwords=stopwords0+stopwords1+stopwords2+stopwords3+stopwords4+stopwords5
print(stopwords)
```

Figure 4. Python language to stop word processing code

### *4.3 Determination of the Number of Topics*

In order to determine the number of topics in the text set, this paper uses the evaluation index consistency score in the statistical language model to determine the optimal number of topics. The formula for consistency score is described as follows[9]:

$$\text{coherence(V)}=\sum_{(v_i,v_j)\epsilon V} score(v_i, v_j, \epsilon) \quad (3)$$

$$\text{score}(v_i, v_j, \epsilon)=log\, p[(v_i, v_j) + \epsilon/p(v_i)\, p(v_j)] \quad (4)$$

In formulas (3) and (4), the probability of each word appearing in the text set, and N is the number of all words appearing in the text set. The consistency score is calculated by the co-occurrence frequency of the words in the sliding window, which increases with the increase of sentence similarity, so the higher the consistency score is, the better[10].

According to formulas (3) and (4), this paper uses Python to write a consistency score program for the theme of LDA model to calculate the value of the consistency score under different subject numbers, and display it in the form of a broken line. 5 is shown. In the Gibbs sampling algorithm, the number of topics is selected, the hyperparameter is set to 50/K, the hyperparameter is set to 0.01, the number of sampling iterations is set to 1000, and multiple clustering experiments are performed using different subject numbers. From Figure 5, the number of topics with the highest consistency score is 20, so the optimal number of topics is determined to be 20.

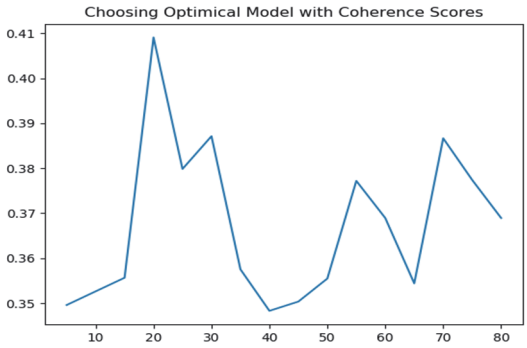


Figure 5. topic consistency score

### *4.4 LDA Model Theme Visualization*

Use the built LDA topic model for topic mining, and determine the best number of topics by consistency score of 20. The presentation of the theme visualization mainly includes: topic-vocabulary dynamic visualization map, sub-list of related information about each topic of the article, and topic category-the article list with the highest contribution rate of the topic.

Use Python's pyLDAvis package to work with jupyter notebook to visualize the topic-vocabulary dynamics, as shown in Figure 6.

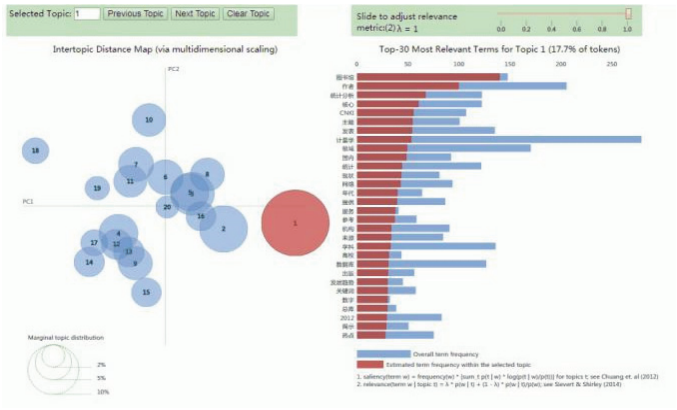


Figure 6. Topic-Dynamic Visualization of Vocabulary (Topic 1)

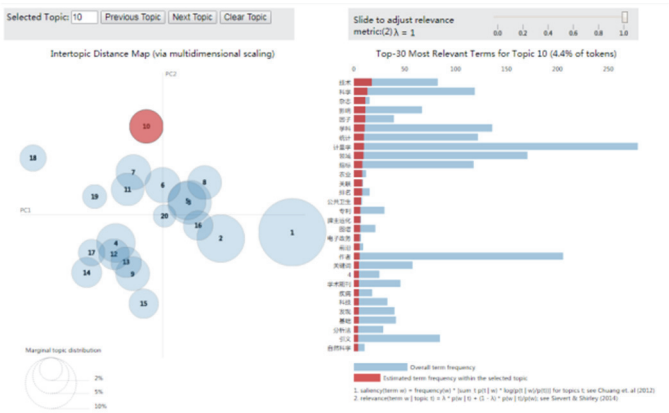


Figure 7. Topic-Dynamic Visualization of Vocabulary (Topic 10)

Figures 6 and 7 are the subject-vocabulary visualization maps for Topic 1 and Topic 10, respectively. The bubble on the left side of the map is the theme distribution, and the bubble represents a theme. The larger the bubble is, the more vocabulary belongs to the theme. The right side will dynamically display the vocabulary of the corresponding theme. The blue bar on the right represents the frequency of vocabulary occurrence under all themes. Indicates how often the vocabulary appears under this topic. Different bubbles represent different themes, and corresponding keywords and words are presented in descending order of relevance. The dynamic visualization of the subject-vocabulary not only shows the vocabulary with the highest contribution rate under each topic, but also the topic of each vocabulary contribution.

Table 1. Article Topic - Vocabulary - Source Information Schedule (Top 10)

| Document_No | Dominant_Topic | Topic_Perc_ Contrib | Keywords | Text |
|---|---|---|---|---|
| 0 | 5 | 0.451 | Nursing, fund, database, metrology, author, funding, clinical, trend, hospital, region | Research status of hospital management in foreign countries:a bibliometric analysis |
| 1 | 9 | 0.9824 | Author, hotspot, keyword, institution, visualization, education, map, CNKI, network, data | A quantitative analysis of the research on college counselors'professional ability based on CNKI database |
| 2 | 2 | 0.8228 | Disease, database, treatment, patient, metrology, clinical, infection, 2016 | Bibliometric analysis on adolescent myopia prevention and control |

| | | | | |
|---|---|---|---|---|
| 3 | 14 | 0.9623 | International, Hotspot, Institution, Web of Science, Plant, Science | Evolution and Research of International Health Promotion:Bibliometric Analysis and Visualization Based on CiteSpace V |
| 4 | 14 | 0.9927 | International, Hotspot, Institution, Web of Science, Plant, Science | A bibliometric appraisal of research on Artemisia from 1986-2016 |
| 5 | 13 | 0.6224 | College, foundation, moxibustion, science, innovation, resources, training, culture, mathematics, team | Bibliometric Analysis on the Reaearch Hotspots of Blockchain in China |
| 6 | 7 | 0.7456 | Reading, Keywords, Heavy metal, Medical, Author, Institution, Database, Pollution, Humanities | A Bibliometrical Analysis of the Exploration Research of Simultaneous Adsorption of Organic Pollutants and Heavy Metal lons |
| 7 | 14 | 0.6722 | International, Hotspot, Institution, Web of Science, Plant, Science | Biblimetrics Analysis on the Trend of International Earth Sciences in Recent 5 years |
| 8 | 9 | 0.4059 | Author, hotspot, keyword, institution, visualization, education, map, CNKI, network, data | Comparisons of Micro Government at Home and Abroad Based on Bibliometric and Visualization |
| 9 | 9 | 0.5144 | Author, hotspot, keyword, institution, visualization, education, map, CNKI, network, data | Retrospection of Hotspot in Educational Management Research in china |

Table 1 is one of the applications of the LDA model, which is to determine the theme of each document.Python builds one containing: Document_No (article number), Dominant_Topic (the subject of the article), Topic_Perc_Contrib (the topic's contribution rate), and Keywords (the subject) ), Text (Article source) List of several elements, by calling the function of the LDA model, generate a predefined list and save it as EXCEL format and save it. The actual generated EXCEL document contains information on the 20 subject papers, the subject of the article, the topic contribution rate, the subject words and the article source of the 857 articles compiled in 2017. Table 1 captures the first 10 articles presented.

As can be seen from Table 1, the articles with the serial numbers 3, 4, and 7 are grouped into the topic 14. The keywords of the class are: international, database, hotspot, institution, Web of Science, plant, science, meaning that these keywords appear together. The probabilities in these articles are the greatest, and the subject contribution rates are 0.9623, 0.9927, and 0.6722, respectively.

Table 2. Topic-Vocabulary-Source of Contributions with the highest contribution rate

| Topic _Num | Topic_Perc _Contrib | Keywords | Text |
|---|---|---|---|
| 0 | 0.9944 | Theme, metrology, community sports, science, hotspots, CSSCI, library, article, author | 2005-2015，our country community sports research analysis of the literature metrology |
| 1 | 0.997 | Team, coronary heart disease, author, library, Chinese medicine, Chinese medicine, trend, database | Bibometrics analysis of status quo of nursing research of colostomy irrigation in China from 2005 to 2015 |
| 3 | 0.9927 | Service, library, agriculture, scholar, | Bibliometric analysis in agricultural |

| | | | |
|---|---|---|---|
| | | patent, culture, science, metrology, theory | Internet of Thing based in Citespace |
| 4 | 0.9898 | Innovation, metrology, cooperation, social sciences, humanities, assessment, author, journal, synergy, literacy | Analysis of Literature Measurement and Co-word Analysis Research Based on I-U-R Collaborative Innovation in China |
| 6 | 0.9948 | Technology, policy, social management, wisdom, knowledge flow, innovation, governance | From social manegement to social governance:A quantitative analysis of knowledge flow and topic evolution |
| 8 | 0.995599985 | Author, metrology, institution, cooperation, characteristics, evaluation, service, discovery, theme | Econo metric Analysts of author collaboration and research hotspots of knowledge service domain |
| 9 | 0.991999984 | Author,hotspot, keyword,institution,visualization,education,atlas,CNKI,network,data | Hotspots and Frontier Themes of Teacher Education Research in China-Based on Bibliometric Analysis of CSSCI Database Papers from 1998 to 2015 |
| 12 | 0.996100008 | Institution, SCI, author, Web of Science, collaboration, database, metrology | A bibliometric analysis of SCI papers published by a medical University based on the Date from 2010-2014 |
| 13 | 0.988799989 | College, foundation, science, innovation, resources, training, culture, mathematics, team | A Biblimetrics Analysis on the Studies of Mathematical Culture in China |
| 16 | 0.995899975 | Evaluation, indicator, data, metrology, influence, citation, science, Altmetrics | Research on Altmetrics Index of Highly Cited Papers based on Article-level metrics |
| 19 | 0.991900027 | Disciplinary structure, discovery, characteristics, keywords, co-words, knowledge map | Research on Disciplinary Structure and Evolutionary Track based on Multidimensional Metrolodh |

Merge the topics of subject 1、2、5、7、10、11、15、17、18 which refer to medical vocabulary,such as drug therapy, clinical infection, hospital care, medical institutions, liver orthopedics, glaucoma, medicine, chronic gastritis, coronary heart disease etc,and merge topics 3 and 14 containing agriculture and plants, so the remaining 11 topics for metrology in 2017.

Topic 0: The application of scientometrics in the field of sports, using CSSCI data for the analysis of topic hotspots; the highest contribution rate is Quan Huifang using literature, bibliometrics, mathematical statistics, comparative analysis, logical reasoning, etc. to CSSCI Sports The source of the journal is a data source, and a bibliometric analysis of the community sports direction in China from 2005 to 2015.

Topic 1: Econometric analysis in the medical field, including author, team and trend analysis; the highest contribution rate is Xue Dongqun et al. using bibliometric analysis, 2005-2015 Chinese biomedical literature database, Chinese journal full-text database, VIP and The literature related to "colostomy lavage care" in the database was analyzed and systematically analyzed, and the development trend of colostomy lavage nursing research in China was systematically analyzed to provide evidence-based evidence for clinical practice.

Topic 3: The application of scientometrics in agriculture is based on patent data; the highest contribution rate is Guo Xiangyun and others using CNKI's 2010-2016 patent data as the research object, using Cite Space to analyze from the literature and social network. From the perspective of the agricultural Internet of Things research team, keyword characteristics were analyzed.

Topic 4: The application of scientometrics in the field of humanities and social sciences; the highest contribution rate is Chuanwangwang's method based on bibliometric analysis and co-word analysis, which is produced in the Chinese Social Science Citation Index (CSSCI) source journals collected in the past 10 years. 1407 papers with the theme of academic research and collaborative innovation are analyzed and researched. Through the distribution of journals in statistical literature, high-yield institutions and high-frequency keywords and co-word matrix construction of the literature, factor analysis, cluster analysis, strategic coordinates and other methods are used for

domestic Hotspot analysis and framework classification were carried out in the field of collaborative innovation.

Topic 6: Applied metrology methods for social policy research; the highest contribution rate is Li Jinxi's CNKI database as a data source, starting with the results of social management and social governance research literature, using scientometric methods, with Bib excel and Cite Space Related functions, quantitative research on the retrieved documents.

Topic 8: Thematic discovery of the metrology method and the cooperation of the authors; the highest contribution rate is Gao Yang's analysis of the authors and cooperation in the field of knowledge services, including the distribution of journals in the field, the number of authors added, the number of posts, High-yield researchers, source regions and institutions, cooperation and co-production rates, types of cooperation, and author cooperation networks.

Topic 9: Visualization analysis of CNKI's educational data and related hotspot research; the highest contribution rate is Wang Hao's analysis of the growth trend of literature in the field of teacher education in China, the distribution of journals, the author's publication volume, cooperation, and team growth. The degree of author activity, as well as the distribution of authors and sources of cited literature, and the research knowledge base and frontier topics of ground-breaking literature, high-frequency citation analysis, citation literature clustering, and key node literature.

Topic 12: Analysis of author cooperation and institutional cooperation on WOS and SCI data; the highest contribution rate is based on the comprehensive search of web of science database, JCR and ESI, and the collection is completed with a medical university. The first or correspondent author of the unit published the SCI paper, and analyzed the SCI paper high frequency author, high frequency cooperation organization and high frequency keyword co-occurrence in the form of science and technology map.

Theme 13: Quantitative analysis of scholars and papers in college mathematics disciplines; the highest contribution rate is the number of mathematics cultural papers published by Yan Changgen et al. in core journals, the author's source, the subject of interest, and the subject of research. Systematic analysis and quantitative analysis were carried out in terms of research methods and other aspects.

Topic 16: Using scientometrics for evaluation analysis (influence) and citation analysis; the highest contribution rate is Huang Xiao's top papers in the field of “social science and synthesis”, which is obtained through ESI database and Altmetric.com. The article details of the social science, comprehensive "field literature", cited frequency, Altmetrics indicators and other data. The high-cited papers construct the influence index system and stratify the Altmetrics index. Then the feature analysis and usability analysis of the Altmetrics index reveals that the Altmetrics score can comprehensively reflect the social influence of the highly cited papers. The Altmetrics score is mainly recommended. , citation, discussion of the typical indicators in the three types of indicators, the number of news mentions, the number of blog mentions, the number of tweets mentioned, and from the statistical analysis results and usability indicators, policy mentions, Facebook mentions, wikis Encyclopedia citations, Google+ mentions, and readings in Mendeley are also worthy of attention.

Topic 19: Analysis of the evolution of disciplinary structure based on co-word analysis; the highest contribution rate is Li Huizhen's analysis of key words, topics, authors, literature and other knowledge units in the core journals of the past five years in the field of library and information, using content analysis Various co-word analysis, cluster analysis, co-citation and cooperative network analysis, etc., analyze the disciplinary structure and discipline evolution from the micro, meso and macro perspectives.

## 5. Conclusion

Based on LDA's topic mining model, this paper can cluster domain topics in depth and visualize the clustering topics. The 2017 metrology data is used as an example to verify. Through data analysis, it can be confirmed that: 1. Combining the LDA mining model with the Python-based pyLDAvis visualization method can not only dynamically render the theme and corresponding vocabulary in both directions, but also the topic saliency and vocabulary. The presentation of contribution rate is also clear to achieve a better user experience; 2. The topic-vocabulary-source list can present all the documents in the form of topic categories, topic contribution rates, keywords and document source titles, subject-vocabulary- The highest contribution rate source table presents each topic with the subject category, the highest contribution rate, the highest contribution rate document source title and keywords.

Inadequacies: 1. In terms of data, the lack of data sources in the data source has led to a large proportion of the sources of literature with the highest contribution rate. Data processing can be considered by machine learning methods based on part-of-speech tagging and training, more intelligent use of useless words (such as verbs,

quantifiers, conjunctions, etc.); 2. About model optimization, through the consistency score to select the optimal number of topics, there are some optimization methods that can improve the consistency score of the model.

**References**

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J Machine Learning Research Archive, 3,* 993-1022.

Chen, J., & Lu, Q. (2015). Visualization analysis of topic mining research. *Library Science Research, 14,* 2-11. https://doi.org/10.15941/j.cnki.issn1001-0424.2015.14.001

Cui, J., Du, W., & Guan, Y., et al. (2017). Research on the Evolution of Micro-blog User Information Personalized Recommendation Model Based on LDA. *Information Science, 35*(8), 3-10. https://doi.org/10.13833/j.cnki.is.2017.08.032

Fen, J., & Zhang, Y. (2017). Research on the Method of Detecting and Analyzing Scientific Fronts Based on LDA and Ontology. *Information Studies: Theory & Application, 40*(8), 49-54. https://doi.org/10.16353/j.cnki.1000-7490.2017.08.009

Guan, P., Wang, Y., & Fu, Z. (2016). Effect Analysis of Scientific Literature Topic Extraction Based on LDA Topic Model with Different Corpus. *Library and Information Service, 2,* 112-121. https://doi.org/10.13266/j.issn.0252-3116.2016.02.018

Guo, L., Li, Y., & Mu, D. et al. (2016). A LDA Model Based Topic Detection Method. *Journal of Northwestern Polytechnical University, 34*(4), 698-702.

Hu, J., & Chen, G. (2014). Mining and Evolution of Centent Topics Based on Dynamic LDA. *Library and Information Work, 58*(2), 138-142. https://doi.org/10.13266/j.issn.0252-3116.2014.02.023

Li, X. D., Zhang, J., & Yuan, M. (2014). On Topic Evolution of a Scientific Journal Based on LDA Model. *Journal of Intelligence, 7*, 115-121.

Lin, F., Xiahou, J., & Xu, Z. (2016). TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model. *Multimedia Tools & Applications, 75*(22), 1-30. https://doi.org/10.1007/s11042-016-3363-9

Liu, J. H., & Cui, J. M. (2016). Domain-based hot research topic mining based on VOS viewer. *Intelligence Exploration, 2,* 13-16.

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. Pergamon Press, Inc. https://doi.org/10.1016/j.eswa.2017.03.020

Ruan, G., & Xia L. (2017). Retrieval Results Clustering Application Research Based on LDA. *Journal of Intelligence, 36*(3), 179-184.

Shi, Q. K. (2013). Document tipic mening and document static visualization based on LDA. Master's thesis, Guangxi University, Guangxi province, China.

Song, Kai, Li., Xiu, xia., & Zhao, S., et al. (2017). International Knowledge Flow Analysis Based on LDA Model. *Journal of Intelligence, 36*(6), 55-60.

Stevens, K., Kegelmeyer, P., & Andrzejewski, D. et al. (2012). Exploring topic coherence over many models and many topics//Conference on Empirical Methods in Natural Language Processing.

Tang, X. B., & Xiang, K. (2014). Hotspot Mining Based on LDA Model and Microblog Heat. *Library and Information Service, 58*(5), 58-63. https://doi.org/10.13266/j.issn.0252-3116.2014.05.010

Theis, L., Aron, van. Den. Oord & Bethge, M. (2016). A note on the evaluation of generative models. Computer Science.

Wang, L. (2015). Research on Text Mining Based on Topic Model (Master's thesis, Dalian University of Technology, Dalian city, China).

Wang, P., Gao, Ch., & Chen, X. M. (2015). Research on LDA Model Based on Text Clustering. *Information science, 1,* 63-68. https://doi.org/10.13833/j.cnki.is.2015.01.013

Wang, W., Zhou, Y. M., & Yang, A. (2017). Method of Sentiment Analysis for Comment Texts Based on LDA. *Journal of Data Acquisition and Processing, 32*(3), 629-635. https://doi.org/10.16337/j.1004-9037.2017.03.023

Yang, Ch., Zhu, D. H., & Wang, X. F., et al. (2017). Technical Topic Analysis in Patents: SAO-based LDA

Modeling. *Library and Information Service, 3,* 86-96. https://doi.org/10.13266/j.issn.0252-3116.2017.03.012

Zhang, L. (2016). Research on Tagging Recommendation Method Based on LDA Topic Model. *Journal of Modern Information, 36*(2), 53-56.