Scientific Journal of Information Engineering
May 2017, Volume 7, Issue 1, PP.15-26

Research on the Development of Text Mining Technology based on Bibliometrics and Knowledge Map Visualization*

Qiuxue Xu[1], Na Niu[1], Yongmin Quan[1], Zhezhi Jin[1†]

1. College of Science, Yanbian University, Yanji 133002, China

†Email: jinzhezhi@sina.com

Abstract

Text mining is a process of analyzing the text data, extracting the potential information, and discovering the text knowledge. This article uses the literature measurement analysis method, with the help of Bicomb's bibliometric function, Ucinet's social network analysis function, statistical analysis function of SPSS and visualization analysis function of Netdraw、Citespace, From the aspects of the distribution of literature, the distribution of journals, the cooperation of high yield authors, institutional cooperation, research focus and research frontier, the present situation of the research of text mining is presented, which provides a reference for the theory and practice of text mining in China.

Keywords: Text Mining; Bibliometric; Knowledge Map; Research Hotspot

1 INTRODUCTION

In the real world, the information is stored in the form of text, such as news pages, patent information, medical records and the Web page. The information is huge, unstructured, and people or computers have difficulty accessing useful information directly. In the face of these huge information, people are eager to make reasonable and effective analysis, and the appearance of text mining provides great technical means. Text mining is an emerging knowledge discovery technology which has high application value in the field of business. The text mining was first proposed by Feldman[1] in 1995, and the text mining is derived from data mining, which refers to the process of extracting potential, undiscovered and useful knowledge from free texts[2].The text mining involves a wide range of fields, covering key technologies such as data mining, information technology, content analysis, machine language learning, statistics, visualization and other key technologies.

As a multidisciplinary cross research field, there are very high research value in the field of multiple directions, so the current situation of research in the field of Chinese text mining research appears very necessary. Through information measurement and visualization analysis of core journal literature of Chinese text mining, which was collected by CNKI from 1998 to 2016,to show the outline of the text mining research, grasp the text mining research trend, clear of the development of the study of text mining, make broad scholars more intuitive understanding of the Chinese text mining knowledge structure and characteristics of the study.

2 DATA SOURCE AND RESEARCH METHODS

2.1 Data Source

The knowledge content of the form of text is usually presented in the evolutionary cycle. By studying the evolution rule, it is found that the research hot spot and trend in the scientific literature are a hot research direction in the field of text mining in recent years. The source of this article is the CNKI database. In CNKI, the topic of "text mining" is

*Fund support：By the Jilin Provincial Department of Education, "thirteen five" scientific and technological research projects to support funding (Kyrgyzstan teach the word [2016] No. 248)

retrieved, the source category is "core journal" and "CSSCI". The literature time is 1998-2016 and the retrieval time is March 2, 2017.In order to improve the accuracy of the literature, the publication requirements、conference notice、news report and the unmarked key words were eliminated, Finally 869 articles were received.

2.2 Research Methods

Bibliometric analysis is a method to use the the relevant knowledge and theories of bibliometrics, with the help of the quantitative characteristics of the literature, describe and evaluate the current status of a subject area by mathematical and statistical methods[3].This study adopts the method of bibliometric analysis, and analyzes the relevant literatures from four aspects, such as the time distribution of literatures, the distribution of magazines, the cooperation of high-yielding authors and the cooperation of institutions.

Knowledge map is a graph showing the development process and structure relation of a domain knowledge, and has the dual characteristics of "graph" and "spectrum"[4].At present, the common analysis method is to construct the keyword co-word matrix by means of using coword analysis to represent the contents of the research field. The principle of co-word analysis is statistics the number of occurrences of noun phrases together in the same literature, based on the above analysis results, cluster analysis of these words is carried out, revealing the relationship between these words and analyzing the structural changes of subjects represented by these words[5].

The data collection and statistics of this paper are based on the BICOMB developed by Cui Lei in the department of medical information of China medical university, The system provides the bibliometric analysis function, which can extract and count the information of the paper, form and output the co-occurrence matrix. The cluster analysis of this paper applies social science statistical software SPSS, which can be used to excavate data statistics.Social network analysis adopts UCINET, the software is a multi-functional, comprehensive social network analysis software,it integrates the analysis of one-dimensional and two-dimensional data Netdraw software, can be graphically visualized to display the data, while also providing a large number of data management and conversion tools. Citespace,a visualization tool developed by Dr. Chen Chaomei of drexel university of America, is able to construct co-occurrence map, co-directed and time-zone views, which are characterized by the exploration of trends in the frontiers of the research field.

3 THE OVERALL SITUATION OF TEXT MINING RESEARCH IN CHINA

3.1 Age Distribution of Literature

The analysis of the age distribution of the literature reflects the research development speed and level of the subject to some extent[6].The data was imported into BICOMB and Excel software for the annual distribution of literature statistics, the specific results are shown in figure 1.

As can be seen from Figure 1, before 2005, the volume of publications is not only small, but also the increment was relatively stable. This period can be regarded as the initial stage of text mining research. From 2006 to 2011, the number of papers has been a certain growth, the annual amount of papers is about twice as much as the previous. This is an important period of Chinese text mining research, indicating that Chinese scholars have a certain understanding of text mining and gradually pay attention to deeper areas to explore. Since 2012, the number of text mining literature has remained at a high level. This is because with the technology research of various disciplines on Text Mining concerns, the depth of the published papers continue to increase and quality significantly improved. This means that text mining as a new research field is becoming a research hotspot, more and more scholars are concerned about and apply this technology to carry out relevant academic research work.

3.2 Journal Distribution of Literature

Through the analysis of journals source of text Mining, not only helps to understand the distribution characteristics of text mining research, but also finds out the core journals in this field[7].The 869 searched articles were arranged in descending order according to the amount of the literature, in which the top 10 journals were shown in Figure 2.According to Brad Ford's law[8]:If the scientific journals are sorted according to the quantity issued, the journals

can be divided into the core area, the relevant area and the non relevant area. At this point, the number of articles in each area is almost the same, but the number of journals shows the relationship between $1:n:n^2$. According to the above law, the amount of journals will be divided into roughly 3 areas, the specific data are shown in table 1.

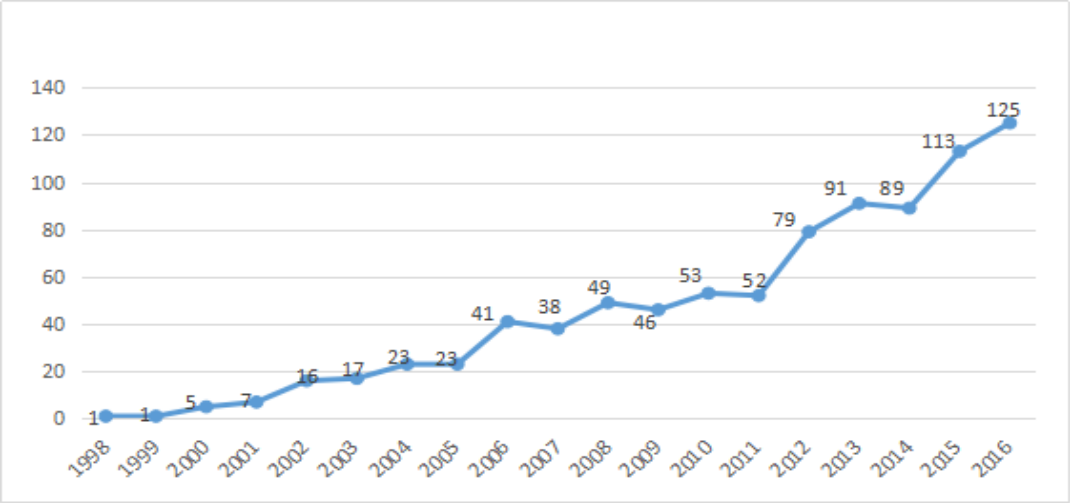


FIG. 1 CHINESE TEXT MINING STUDIES DISTRIBUTION OF AGE DISTRIBUTION

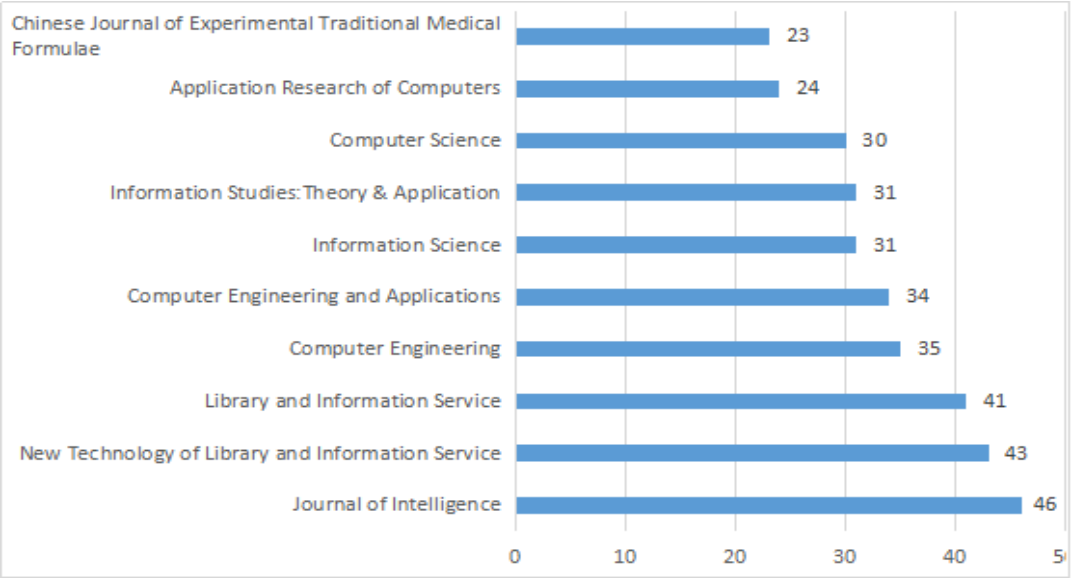


FIG. 2 STATISTICS OF JOURNAL DISTRIBUTION IN TEXT MINING (TOP 10)

TABLE 1 JOURNAL DISTRIBUTION LIST TABLE

area	division standard	The number of journals	The number of articles	Proportion of the total number of papers(%)
core area	>24	8	291	33.49
relevant area	4~24	39	334	38.43
non relevant area	<4	177	244	28

As can be seen from table 1, the number of journals in the three areas is 8:39:177,and the n is about 4.7. Therefore, it is basically consistent with Brad Ford's law. That is to say, the journals distribution in the field of text mining has formed a stable core journal group. Most of these journals are library science, information science journals, or computers. The largest amount of papers is "intelligence magazine", which contains 46 related papers, accounting for 5.2934% of the total. Among the 47 journals in the core area and related area, 21 journals are from the CSSCI journal, 41 journals are the Chinese core journal. The papers published in these journals have high citation frequency and high reprint rate, also have high innovation and academic level, which has effectively promoted the development of Chinese text mining research.

3.3 Prolific Authors Collaboration Analysis

The authors of text mining study group determines the development level of the field, The core author team of this field can be determined by the statistics and analysis of the author's team, which has guiding significance to grasp the current situation and development trend prediction research in this field. According to Price law (the number of authors who publish the most papers is 49, the square root of 49 times 0.749, which is 5.243, and the number is 5),authors who publish five or more papers can be considered prolific authors. According to statistics, the output of the prolific authors accounted for 43.3% of all papers. According to price's law, the proportion of prolific authors in all the papers is not up to 50%, which is to say, the field of text mining has not formed the core authors group. At the same time, selects the author whose quantity is more than or equal to 5 to construct the co-occurrence matrix. After Ucinet conversion, the data file is imported into NetDraw to obtain the prolific author cooperation network map (see Figure 3) and display according to the center size.Each node in the graph represents an author, and the size of the node is related to the frequency of the author's published papers. The thickness of the connection between nodes represents the number of times of collaboration between authors.

We have observed three academic groups through the Cooperative network map of prolific authors (Figure 3):

(1)With Zheng Guang,Lv Aiping, Jiang Miao belong to China academy of traditional Chinese medicine clinical foundation of traditional Chinese medicine such as institute of scholars constituted a large academic research team, and Guo Hongtao of the Shanghai University of Traditional Chinese Medicine, Lin Hongfei of Northeastern University and Huang Lucheng of Beijing University of Technology have close cooperation and communication with the team.The team uses text mining technology to explore the medication rules of various diseases, the characteristics of drugs and the treatment rules of diseases with proprietary Chinese medicine and Western Medicine.

(2)Qu Peng, Wang Huilin and Zhang Chengzhi from China Information Institute of science and technology constitute a small academic team. The term extraction in patent information service and the problem of multilingual text clustering.

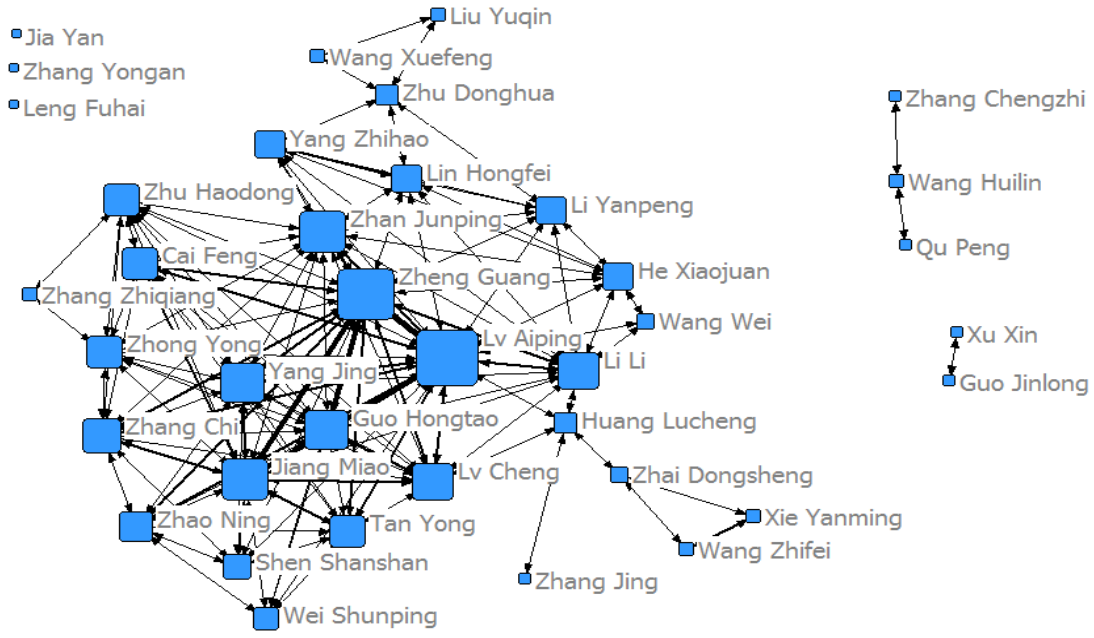


FIG. 3 COOPERATIVE NETWORK MAP OF PROLIFIC AUTHORS

(3)Guo Jinlong and Xu Xin of East China Normal University. Their academic research focuses on the application of text mining in the study of Human and Social Sciences.

Therefore, the core author group in the field of text mining only have one large academic team, and the rest is based on the cooperation between teachers and students within an organization. So China needs to further strengthen cooperation in the field of text mining research, through the integration of resources, organizational interdisciplinary

cooperation, give full play to their respective advantages, realize the sharing of knowledge, equipment and other resources, make the research in the field of text mining breakthrough, higher output level results.

3.4 Institutional Collaboration Analysis

The number of scientific papers produced by the research institution indicates its research strength in one respect, and the more it has, the more influence it has in the field. The cooperation between institutions reflects the composition of the research team and power distribution[9].According to the statistics, there are 817 research units, Among them, the China academy of traditional Chinese medicine clinical foundation of traditional Chinese medicine was ranked first with 56 articles. After selecting the research institutions with more than 5 papers to construct the co-occurrence matrix, and then import the data files into NetDraw through Ucinet conversion to obtain the cooperation network map of the core research institutions in the text mining area (Figure 4).

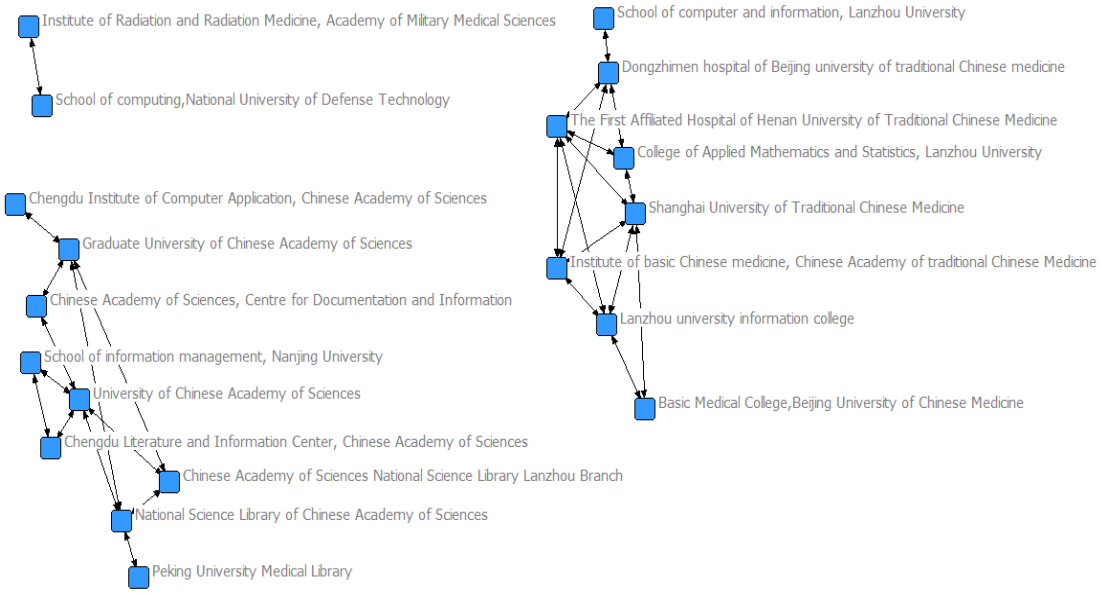


FIG. 4 COOPERATIVE NETWORK MAP OF CORE RESEARCH INSTITUTIONS

Figure 4 shows China's core research institutions and their cooperative relationship. Among them, there are two cooperation groups composed of three scientific research institutions. The first group is composed of 9 institutions, including China University Academy of Sciences, National Science Library of Chinese academy of sciences, the Chengdu Literature and Information Center, of Chinese Academy of Sciences and other. It can be seen that the core of the team is the Chinese Academy of Sciences, and the team members are the different branches of the Chinese Academy of Sciences, different agencies and engineering centers. Chinese Academy of Sciences is the highest academic institution in the China natural sciences, their academic research field is very comprehensive, including not only key technologies and system development of text mining, but also base on the text mining technology for goods comments and patent analysis. The second group is composed of 8 institutions, including the First Affiliated Hospital of Henan University of Traditional Chinese Medicine, Institute of Basic Theory traditional Chinese medicine of Chinese Academy of Traditional Chinese medicine and the Shanghai University of Traditional Chinese Medicine. Their academic research focuses on the use of text mining techniques to explore the characteristics of various diseases and medication rules. Another group is made up of two institutions, there are the Institute of Radiation and Radiation Medicine of Military Academy of Medical Sciences and the computer academy of National Defense University, their academic research focuses on the methods of text mining based on these ontologies.

4 RESEARCH HOTSPOTS AND FRONTIERS

The research hotspot is the focus of research content in a certain period. Bibliometrics generally shows the hotspots in this field through the knowledge map. Key words are the essence of an academic paper. The research topic represented by the key words which have high frequency and high center is often used to identify the research focus

in a research field[10].

4.1 Analysis on Easearch Hotspots

1) Keyword and its Matrix Extraction

In order to reduce the interference of low-frequency words and show the construction of knowledge points in the field of text mining truthfully, this paper uses the word frequency not less than 10 times high-frequency keywords to analyze. Remove 4 ontologies and keywords that are too broad (text mining, data mining, Web text mining, big data). The 28 effective high-frequency words are determined ultimately, frequency was 554 times, accounting for the total keyword frequency of the 14.86%. If add the highest frequency keyword "text mining" 453 times, the 29 keywords accounted for 27% of the total frequency.Therefore, the 28 high-frequency keywords basically represent the research hotspots in the field of text mining, as shown in Table 2.

TAB.2 HIGH FREQUENCY KEYWORD LIST

number	keywords	number	keywords
1	Text clustering	15	Natural language processing
2	Patent analysis	16	Online reviews
3	Text classification	17	Association rules
4	Knowledge discovery	18	bibliometrics
5	Information extraction	19	bioinformatics
6	microblog	20	Feature extraction
7	Information analysis	21	Knowledge management
8	Traditional Chinese medicine	22	Medication regularity
9	ontology	23	syndrome
10	Feature selection	24	Machine learning
11	Vector space model	25	Domain ontology
12	similarity	26	visualization
13	Support vector machine	27	competitive intelligence
14	Data slicing algorithm	28	Information retrieval

2) Cluster Analysis and Multidimensional Scaling Analysis

It is not sufficient to explain the research hotspot of text mining only by word frequency statistics, so we use BICOMB to construct co-word matrix for high-frequency words to form 28 * 28 matrix. Using SPSS to transform co-word matrix into correlation matrix, and then convert to dissimilarity matrix for clustering. Using the method of system clustering in SPSS software, the above dissimilarity matrix is clustered. The method of Ward connection is adopted, and the square Euclidean distance is chosen as the metric, the specific clustering tree is shown in Figure 5.

Multidimensional scaling is designed to use two-dimensional or three-dimensional space distances to represent distances between individuals, and try to carry out the observation and analysis of individual distance and explore the data structure implicit in it and study the similarity relationship between individuals[11].Through the multidimensional scaling analysis, the results of the previous cluster analysis can be corrected. Similarity can be an important condition to distinguish categories, objects in the same category have strong similarities, and the closer they are to the middle, the more important their position is[12].The specific method is used for scale analysis functions in SPSS statistical software, the specific multidimensional scaling maps is shown in Figure 6.

3) Topic Hotspot Identification

Based on the above all the results and actual semantics for each keyword, the high-frequency keywords can be divided into 4 cluster groups, as shown in Table 3. On this basis, the 28 high frequency keywords in the field of text mining are deeply analyzed combined with the specific contents of relevant literatures.

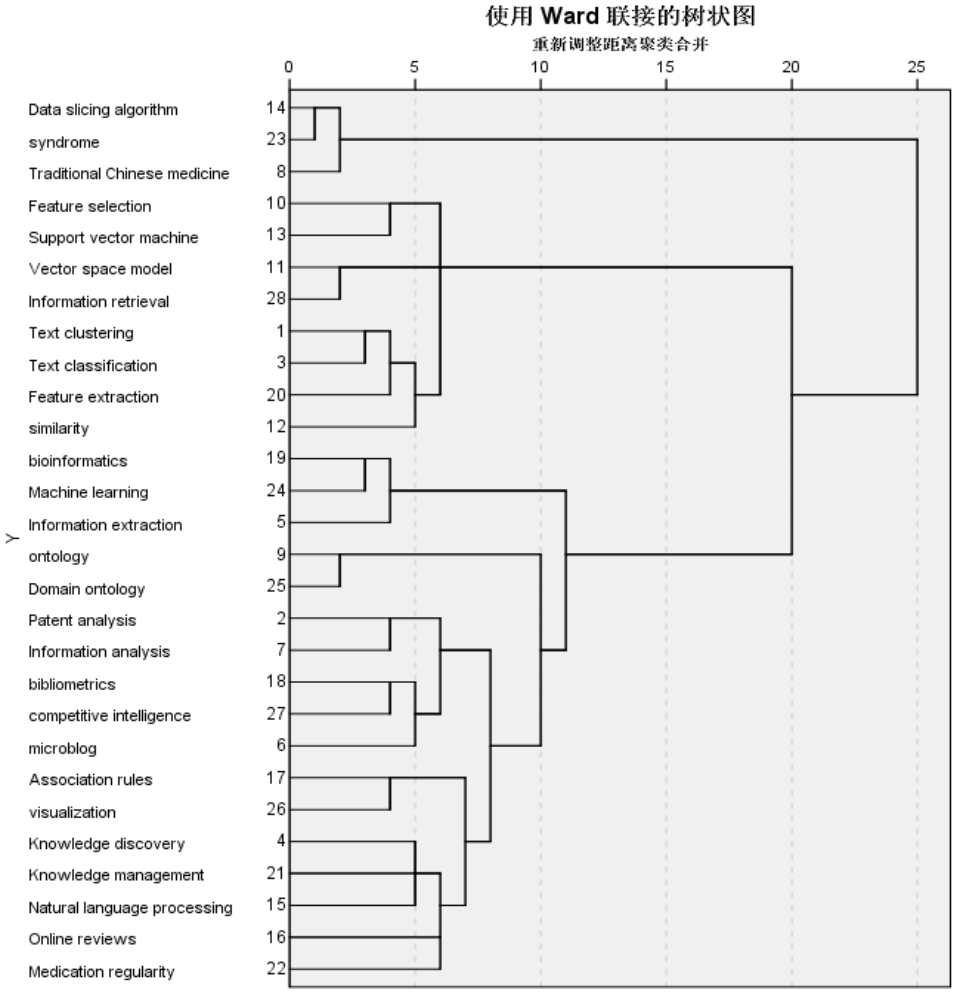


FIG. 5 CLUSTERING TREE STRUCTURE IN TEXT MINING FIELD

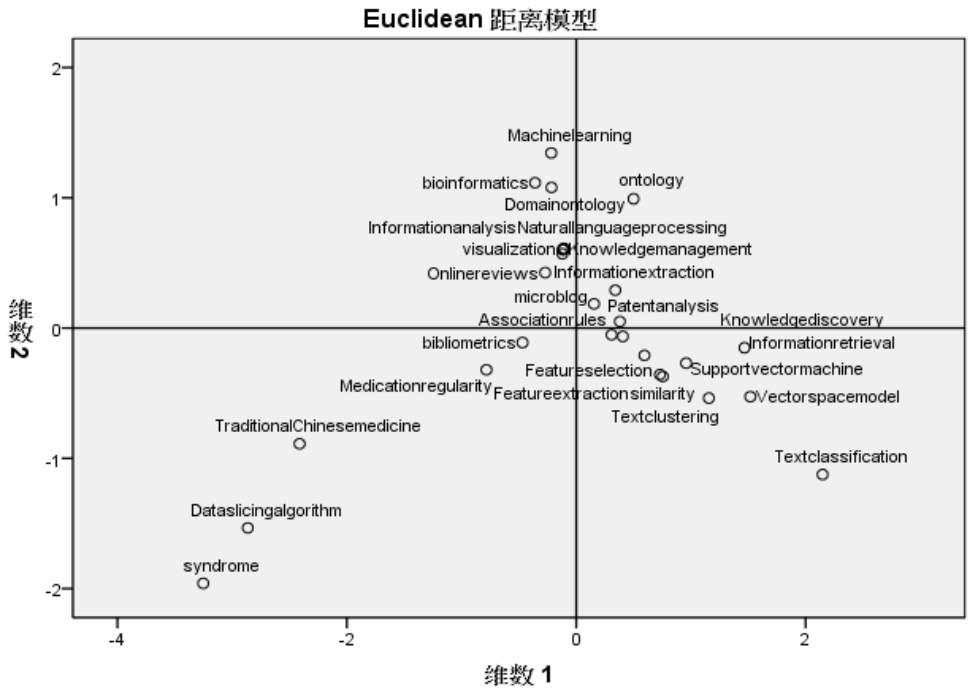


FIG. 6 MULTIDIMENSIONAL SCALING ANALYSIS DIAGRAM IN TEXT MINING FIELD

TAB.3 CLUSTER DIVISION GROUP IN TEXT MINING FIELD

GROUP	KEYWORDS
1	data slicing algorithm、syndrome、traditional Chinese medicine、medication regularity
2	ontology、Domain ontology
3	Natural language processing、Text classification、Support vector machine、Vector space model、Information retrieval、Feature extraction、Feature selection、Text clustering、similarity、Association rules、visualization
4	bioinformatics、Machine learning、Information extraction、Information analysis、Knowledge discovery、Knowledge management、bibliometrics、Patent analysis、competitive intelligence、microblog、Online reviews

According to the results shown in table 4, and the specific analysis of the literature, it can be:

(1)Application of text mining in medical field

A great deal of information about medical history, diagnosis, treatment and so on are contained in medical texts. It is important for the development of medicine to dig out the knowledge hidden in text by using text mining technology. Yang Jing[13] collects the literature data of coronary heart disease in the Chinese biomedical literature (CBM) database, and uses the data hierarchical algorithm to excavate syndromes traditional Chinese medicine and the laws contained in it of coronary heart disease. The study shows that the core drugs of coronary heart disease are consistent with clinical commonly used drugs. Therefore, the text mining technology can provide a technical path for clinical application.

(2)Application of ontology in text mining

The ontology's concept comes from western philosophy, which means to explore the nature of existence and think about the theoretical definition of objective things. The concept of ontology generally includes four levels of meaning[14], conceptualization (abstract the characteristics of objective things and construct suitable models for them); clear(there is a clear explanation of the connection between concepts and concepts), formal(concept can be recognized by computer language), shared (A set of conceptual terms that a professional can understand in the field). Domain ontology describes the relationship between concepts and concepts in the specified domain, as well as the classical theories which are representative, innovative and valuable in the field. The traditional text mining method is to vector space model as the representative takes the vector of the word frequency instead of the text, and the next step is further mining based on this vector. But the method just cut the document into words, ignoring the semantic information in the document, so it is difficult to achieve a higher degree of accuracy. The combination of ontology and text mining can improve the accuracy of the result.

(3)Text processing and mining technology

A complete text mining process includes the following steps: data preprocessing, pattern mining, evaluation and other steps. In general, a text mining process should be integrated with a variety of text processing and mining techniques, such as Chinese word segmentation and feature presentation technology in data preprocessing technology, text classification and clustering, association rules, trend prediction techniques forecasting in the mining analysis and information visualization technology[15].

For Chinese text, vocabulary is the main source of sentences formation. In the process of text mining, first needs to carry on the automatic segmentation and part of speech tagging, this is crucial for improving the accuracy of text interpretation. At present, domestic and foreign researchers have developed some high precision Chinese Natural Language Processing tools, such as computing cloud developed by Harbin Institute of Technology, ICTCLAS system developed by Chinese Academy of Sciences, the segmentation accuracy has reached more than 95%,and various

functions such as speech tagging, dependency syntactic analysis, named entity recognition and semantic role labeling are also used. Association rule is one of the most common applications in text mining. It can help users shorten retrieval time, improve retrieval efficiency, and retrieve the similar documents quickly and efficiently in a large number of texts. In the mining of association rules, the frequency and intensity of rules are expressed by in terms of support and confidence, and users can set the parameters of support and confidence according to their own requirements. Apriori algorithm is a classical algorithm of association rules. Many papers in different fields have applied this algorithm or improved it to obtain the ideal results.

Text categorization is done by machine automation, which classifies text data into several predefined subject categories. The general steps for text classification is to stablish feature set, feature extraction, calculation of feature weight, classification model (classifier) training and classification, evaluation of results and so on. Vector space model is a commonly used text representation model, it takes the keywords in text as the characteristics of text. After certain algorithms are used to reduce the dimensionality of the feature and calculate the weight of the text keyword. Vector space model originated in the text information retrieval ,it made an important contribution to the basic principle of information retrieval and for the subsequent various text information processing operations laid the foundation of mathematical computing. In the actual process, feature extraction is needed to reduce the dimension of feature. The commonly used dimension reduction is feature selection and feature extraction. Feature selection is to remove the feature items with too small weight, the principle of feature extraction is to use the mapping principle (or transform) method to reconstruct the original variable and transform the new features with fewer dimensions. The preprocessing of text clustering and text categorization includes the process of word segmentation, text representation and feature selection. However, text clustering does not require preassigned topic categories and it is more automated. In text classification, text clustering and other text processing technologies, text similarity algorithm has always been a hot research area. The quality of the similarity algorithm has a direct impact on the accuracy of the subsequent classification and clustering. At present, the commonly used algorithms are based on text structure features, keyword matching, semantic similarity degree algorithms, etc.

Text mining and visualization technology are closely linked, text mining technology can show the deep meaning hidden in the text data, and through the visual way to help readers more clearly understand the text.

(4)Application of text mining in other fields

Bioinformatics is an interdisciplinary subject that accompanies the development and accomplishment of the human genome project. It is the problem of predicting and solving genetic related life science by means of computer, statistics and other informatics methods[16].The two popular research themes of bioinformatics are the search for disease-causing genes, gene functions and disease prediction in protein-protein interaction networks. However, a tricky problem is the identification of the biological named entity. The same name can represent different genes/ proteins, while a protein /gene also has multiple names. Therefore, using the information extraction technology in text mining, we can identify and extract the relevant information of protein, gene or disease from the literature summary for the next step of mining.

Text mining is also an important technique in the field of information analysis, which aims at extracting latent and implicit knowledge from unstructured text databases. A high-level text mining system should have automated capabilities, so machine learning methods are employed in almost all systems, which greatly reduces the amount of manual effort. From feature selection to pattern mining and evaluation, machine learning method runs through a complete text mining process. Information extraction enables unstructured free text to be transformed into structured data and stored in a database. The rapid development of information technology in modern society promotes the connection between people and computers more closely, but computers cannot recognize and deal with the language of human normal communication. Natural language has a strong richness and the times, different grammar, semantics and pragmatics can influence the connotation of natural language, and Natural Language Processing is difficult to accurately understand the text information and effective communication. As a result, the exploration and development of Natural Language Processing and related technologies based on a thorough understanding of natural language is now urgently needed to be solved. Information analysis is a scientific method of finding valuable information from information sources and providing services for scientific decision-making. Information retrieval is

actually a matching process. The computer accepts the user's requirements, then finds the relevant information in the database, and finally feedbacks it to the user in an appropriate way.

With the arrival of the knowledge economy, knowledge has become the dominant factor in the production process. Text mining, also called text knowledge discovery, is to analyze the text data so as to dig out the internal knowledge. Searching and digging through comments on web users is a way of knowledge discovery, and it is also a hot research issue. For enterprises, years of operating experience and other tacit knowledge, it can be converted into explicit knowledge through effective knowledge management so as to realize the sharing of knowledge and improve the performance of enterprises.

Literature is a record of scientific research, bibliometric analysis is a method of analyzing scientific literature by measuring some indexes, which is an important means to analyze the structure and trends of disciplines and specialties. Bibliometric analysis makes quantitative statistics on the published information, including the annual distribution of papers and citations, the distribution of authors and agencies, and the analysis of their cooperation, but it only emphasizes numbers and ignores content. The related techniques of text mining, such as text clustering, can be used to analyze the hot spots of literature, some prediction models can also be predicted and verified.

As a record carrier of inventions, patents contain a lot of technical information. The correlation analysis of application text mining can be used to identify the relevant core technologies of patents, find suitable alternatives and even find cooperation opportunities. Patent analysis is widely applied to enterprise competitive intelligence collection, and plays a guiding role in the enterprise strategy formulation. Competitive intelligence is based on facts, and the real-time data for making decision and strategy development, it’s mining method is to understand the industrial analysis of industry participants. From the Web page to extract the real meaningful text information, and carries on the weight and classification, then there is a clear analysis of enterprises and their competitors' advantages and disadvantages .These processed intelligence information is valuable to the enterprises, and it also saves a lot of time for the enterprise.

As a form of social media, microblogs generate huge data, using text mining technology can dig out valuable information. For example, xu fan[17] excavates the content of sina weibo's official weibo account, through the analysis of the comments, find out the content of the user's interest and influence the strength of the analysis. The mining results can help lefeng.com understand user interests and needs, but also provide some suggestions for the enterprise in marketing .

As the network shopping has become an indispensable part in people's daily life, product reviews have also become an important reference for consumers before consuming. At present, The application of text mining technology in network comments analysis reflects in three aspects: information extraction, sentiment analysis and text classification. Online comments on the mining theme is relatively small, and the accuracy of sentiment analysis is not high enough. In addition, the identification of false reviews and the motivation of consumer reviews will be potential research hotspots [18].

It is worth emphasizing that the above 4 categories are not exist in isolation, and they are closely related to each other. In addition, the essence of Co-word analysis is to use the frequency of co-occurrence word to reveal the semantic implicit in the words. Clustering is just a statistical method, which artificially divides into the vocabulary of the clustering into groups to represent the subject research[19]. Sometimes the words that are in the same category just have a strong similarity in numerical proportions, but there is no substantive link on the actual semantics, which caused a great deal of bias the results. So, at the time of final categorization, the actual meaning of the words can be used to achieve a more satisfactory effectt.

4.2 Research Frontier

A new theory that is emerging in the field is called the research frontier. In the Citespace software, keywords co-occurrence study of time zone map can show the hot spot, evolution and research frontier of research from the beginning to the present. Therefore, import all the literature into Citespace, set the parameters and get the front and trending view of the field of text mining, as shown in figure 7.



FIG. 7 TIME ZONE CLUSTERING VIEW OF FRONTIER AND TREND IN TEXT MINING

Time-zone maps are in chronological order from left to right represents the evolution relationship of research frontier, the upper right corner represents a research frontier consisting of a set of recent hotspot keywords, which represents the research trend in this field to a certain extent. Figure 7 shows the hotspot in recent years, by looking up the corresponding papers of these words, we can sum up the research topics that have been paid more attention in recent years, thus infer the trend of text mining research. It is clear from figure 7,study analysis(学习分析), bioinformatics(生物信息学) , library(图书馆), machine learning(机器学习) , topic model(主题模型),social media(社会媒体) and so on are active in the recent study on text mining, are likely to become research hotspots in future research.

5 CONCLUSIONS

This article adopts the method of bibliometrics analysis and knowledge map,and uses bibliometrics function of Bicomb, social network analysis function of Ucinet, s age distribution analysis of the literature, to get the number of text mining is in a rapid growth stage, which means that as a new research field is becoming a hot research topic; through the analysis of journals, the field of text mining has formed its own core journals, which is mainly published in the journal of library science, information science journal or computer; through the analysis of high-yield authors and institutions, the core research team and institutional groups are obtained. With the aid of SPSS statistical analysis function and Citespace visualization analysis function, we can display the research hotspot and research front of text mining research, for the theory and provide the reference for the study and practice of Chinese text mining.

REFERENCES

[1] Ronen Feldman, Ido Dagan. Knowledge Discovery in textual databases (KDT)[J].In Proceedings of the First International Conference on Knowledge Discovery and Data Mining(KDD-95). Montreal, Canada. August 20-21,1995 :112-117.

[2] Zhang Qun.Text mining technology and its application in patent information analysis[J].Modern Information.2006:209-210.

[3] Huang Wei,Chen Yong.Study on the knowledge map of the development track of educational economics in China [J]. Education & Economy, 2010(3) : 68－69.

[4] Chen Yue, Chen Chaomei.Citation space analysis principle and application: Citespace practical guide[M]. Beijing: Science Press,2014:16

[5] Feng Lu, Leng Fuhai.The theoretical progress of Co-word analytical method[J].Journal of the Chinese library, 2006(6):88-92.

[6] Qiu Junping, Zhou Yi, Lu Hong. Metrological analysis and evaluation of intellectual property research in the field of Chinese library in the past ten years[J]. Library Forum,2014(01):1-6.

[7] Hu Dehua,Zhong Lexi. Visual analysis and comparison of subject librarian research at home and abroad[J]. Library theory and practice,2015(06):36-41.

[8] Wallace D P, Van Fleet C, Downs L J. The research core of the knowledge management literature[J]. International Journal of Information Management, 2011, 31(1): 14-20.

[9] Sun Yusheng, Qiu Rongrong, Huang Chuhui. Research progress on personalized service of digital library in China: research strength and knowledge base [J]. Intelligence journal, 2013 (12) : 175-182.

[10] Zhao Rongying,Guo Fengjiao,Zhao Yuehua.Research on mainstream research field and hot frontier research of scientific metrology [J]. Library and Information Service, 2015(02) :66-74.
[11] Zhang Ling,Zhou Chunlei,Kou Guangzeng. The domestic knowledge service research based on common word analysis [J]. Library and Information Service, 2009(12):64-68.
[12] Zhang Han, Wang Xiaoyu, Cui Lei.Co-word analysis and the number of citations are combined to study the development trend of the thematic areas[J]. Information theory and practice, 2007 (3) : 378-380.
[13] Yang Jing. Study on the correspondence regularity of coronary heart disease medicine and syndrome based on text mining method [J]. Chinese Journal of Experimental Traditional Medical Formulae, 2013,19(21) : 319－322.
[14] Guo Jinlong,Xu Xin,Lu Yujie. Application progress of Chinese text mining in the humanities and social sciences [J]. Library and Information Service, 2012, 56 (8) : 10-17
[15] Fensel D.The semantic web and its languages[J]IEEE Intelligent Systems,2000,12:67- 73.
[16] Zou Quan, Lin Chen, Liu Xiaoyan, Guo Maozu. Text mining methods in bioinformatics [J]. Computer engineering and design, 2011, 32 (12) : 75-78
[17] Xu Fan. Research on the focus of enterprise microblog comments based on text mining [J]. Information management and informatics, 2015(8) : 45－48.
[18] Zhang Xuan. Review of online product reviews[J]. Cooperation economy and technology, 2016 (15) : 80-82.
[19] Zhong Weijin. Standardized study of Co-word analysis: contrastive analysis of clustering effects of subject words and keywords [J]. Library and Information Service, 2011, 55 (6) : 114-118.

AUTHORS

[1]**QiuxueXu**(1993-),Mastergraduate,research direction:application statistics.E-mail: 294573414@qq.com.

[2]**ZhezhiJin**(1979-),lecturer,graduatetutor,research direction:information statistics and insurance actuarial.Now, he is a professor in the mathematics, Yanbian University.