DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.2, MAY. 2021

Contrastive analysis in China and abroad on the Evolution of hot topics in the field of digital library based on LDA model

Chunhui Tan, Mengyuan Xiong

School of Information Management, Central China Normal University, Wuhan, China

ABSTRACT

Revealing and comparing the evolution process of hot topics in the field of Digital Library in China and abroad. [Methods]: Taking data in the field of Digital Library from core journals in CKNI and Web of Science from 1990s to 2020, topics are extracted by LDA model and hot topics are selected based on life cycle theory. Topic evolution paths are generated to contrast evolution of hot topics between home and abroad which are grouped into dimensions of technology and application. It fails to analyze the lagging performance and reasons of research hot topics in the field of Digital Library at home and abroad. In technological dimension of Digital Library, the research content in China lags behind that at abroad. In terms of application dimension, Chinese application tends to focus on social sciences, while application at abroad tends to focus on natural sciences. The evolution of overall research focus is U-shaped, which gradually shifted from technological research to application research, and now turn back to technological dimension. Nowadays, there are also many emerging topics combined with big data technology.

KEYWORDS

LDA Model; Topic Life cycle; Topic Evolution; Digital Library; Hot Topic

1 Introduction

Journal literature is an important carrier of scientific research achievements, especially core journal papers. They not only have high academic and application value, but also reflect the scientific research development level in a certain geographical range and academic field (Wang et al., 2013). Journal papers have a large number and high content value, and contain unique hot spots and frontier research information in the field of science and technology. However, with the information expansion brought by big data, the total amount of journal papers is increasing. In order to objectively and comprehensively reveal and understand the research status of science and technology, how to efficiently identify potential frontier hot topics, and track their evolution path and provide visualization from massive literature has become an important field of Information Science research in recent years.

There are many research results about the methods of Topic Evolution Analysis in China and abroad, which can be divided into three categories.

The first one is topic evolution analysis based on word frequency statistics. Early Kleinberg (2003) proposed in the early stage that keywords representing hot topics in literature can be

mined by analyzing the characteristics of word frequency distribution. Qiu and Lv (2013) used bibliometric theory to analyze the high-frequency keywords and professional terms in the international high impact journals, and tracked the research hotspots and frontiers in the domestic library and information field. Zhang and Ma (2007) further identified the research hotspots in the field of knowledge management in China and abroad through word frequency analysis, and analyzed the research directions, application methods and related disciplines. These methods are convenient to reveal the research hotspot of a certain discipline at home and abroad, but it is not enough to use word frequency as the criterion to identify the topic.

The second type is the topic evolution analysis based on network community. The method was first proposed by Girvan and Newman (2002). Wallace et al. (2009) applied community discovery in complex networks to document based topic recognition.

This kind of method based on network relationship can identify keywords with higher credibility, but it will assimilate the identified keywords, and there is no weight distinction, which leads to the inability to express the topic intensity, and the theme content is not reasonably divided.

The third type is topic evolution analysis based on topic model. Commonly used topic model is Latent Dirichlet Allocation (LDA) model, which was proposed by Blei et al. (2003). As a text mining method of unsupervised machine learning, LDA model can mine potential topics from initial documents. Since it was proposed, LDA model has attracted the attention of many scholars and has undergone many optimization and expansion, such as Blei and Lafferty (2006) further proposed a dynamic topic model, which divided the object documents into time slices and constructed the LDA model respectively, then constructed the topic model for each time slice document, and then constructed the relationship between different time slices according to certain methods such as similarity calculation, etc., and TOT (Topic Evolution Model) which regards time dimension as the endogenous variable of LDA model, and A-TOT (Author Dynamic Topic Model) which is added by Xu et al. (2014) for the author attribute of TOT model In the later stage. At present, some scholars have tried to use this method in the comparative study of a certain discipline and abroad.

To sum up, many scholars have carried out topic evolution analysis based on specific disciplines to explore frontier and hot research topics, and reveal the research hotspots or front and development direction of the field. For the comparative study of a certain discipline between China and abroad, most of them are still based on word frequency and co-word analysis methods, and a few methods using topic model are only based on topic recognition. However, there is a lack of research on the content and characteristics of topic evolution and visualization.

With the development of information technology, the amount of information that needs to be stored and disseminated is increasing, and the types and forms of information are becoming more and more abundant. Obviously, the traditional library mechanism can't meet these needs. Therefore, people put forward the idea of digital library. Digital library is development of traditional library in the information age. It not only contains the functions of traditional library, providing corresponding services to the public, but also integrates many other information resources (such as museums, archives, etc.) to provide comprehensive public information access services. Digital library is a storage of electronic information, which can store a large number of various forms of information. Users can easily access and obtain stored information through the network without geographical restrictions. As soon as the concept of "Digital Library" was put forward, before the government and relevant commer-

cial institutions began to put it into practice and promotion, it has aroused wide attention in the academic circles all over the world, and a lot of research has been carried out and developed well in the aspects of technology and practical application. The digital library provides the necessary information resources for the information superhighway and becomes the main information resource carrier in the knowledge economy society. Digital library will become the public information center and hub in the future society, and the research in this field will be indispensable part of public information facilities.

In the field of information technology related research, China's academic development is booming, but compared with the international research, China's research still lags behind, not only in the number of research journals, but also in the content innovation of research. There are differences in the research level of the same subject between Chinese and English academic language systems. Most of the existing research are focused on the detecting emerging research field of digital library, but there are few comprehensive comparative studies between China and international ones in this field.

In order to solve the above shortcomings of the regional comparative study of subject evolution, this paper takes the field of digital library as an example, topics are extracted by LDA topic model, and divided into different dimensions, so as to analyze the evolution path and intensity change trend, and visually present, so as to reveal the differences and gap in Chinese and international topics in this field more comprehensively, dynamically and objectively. The technical innovation therefore can be promoted, and application can be expanded.

2 Methodology

2.1 Topics Extracting

With the diversified development of text data processing requirements, there are several commonly used topic extraction technologies, such as TF-IDF(term frequency-inverse document frequency), extracting topic words from the perspective of word frequency and inverse document frequency; Text-rank algorithm based on Web page recommendation system and PageRank which is a unsupervised topic extraction algorithm, etc., but the above methods are more suitable for the situation with more noisy words and single output form, and impossible to reveal the potential relationship between the topic words & texts, or among the topic words. In this paper, considering that data of scientific and technological literature text is in standardized format, LDA is suitably used for topic extraction, which is a text mining method based on bag of words model and unsupervised machine learning. It is different from the discriminant model of predicting conditional probability distribution, potential topics from the initial document can be mined without manual marking in advance. Therefore, the application of LDA in literature content analysis will help to reveal the internal relationship of international studies to a greater extent, which also helps to reveal the internal architecture of field research, and explore the evolution path of field research content by constructing the evolution relationship between topics. Therefore, LDA has a better effect on extracting potential topics of scientific and technological literature.

LDA assumes that a document is composed of multiple topics, and topics are composed of words. Firstly, the length of document generated by Poisson distribution $N \sim Poission(\beta)$, then the Dirichlet distribution $\theta \sim Dir(\alpha)$ of each topic in the document are generated, and there will be a topic $z_{mm} \sim Multinomid(\theta_m)$ for each word in the document. Secondly, the dis-

tribution $\varphi_{zmm} \sim Dir(\beta)$ of each word for topic are likewise generated, z and φ are taken as parameters of a polynomial distribution, by which a word is finally determined. The joint distribution of the whole model can be shown in Formula 1.

$$p(w,z,\theta_m,\varphi_k|\alpha,\beta)=\prod_{n=1}^{N} p(\theta_m|\alpha)p(z_{mn}|\theta_m)p(\varphi_k|\beta)p(\omega_{mn}|\theta_{Zmn}) \quad (1)$$

This paper mainly uses Gibbs sampling algorithm to get the overall distribution of topics and words. As an unsupervised machine learning, three super parameters should be determined in advance, α、β、k (the optimal number of topics), and α、β take general default value (Wei & Croft, 2006). The optimal number of topics K is determined by calculation of perplexity. Perplexity is an effective method to evaluate the effect of linguistic probability model and to assist parameter improvement. It is based on information theory and calculates the uncertainty (information entropy) of probability distribution or model. When it is applied to LDA model, the calculation formula is shown in formula 2.

$$Perplexity(D)=\exp\left|-\frac{\sum_{i=1}^{M}\ln P(d_i)}{\sum_{i=1}^{M}N_i}\right|, \text{ and } P(d_i)=\sum_z P(z,d)=\sum_z P(z)P(d|z) \quad (2)$$

In this paper, the meaning of perplexity is the uncertainty of the subject to which the document belongs, so the more perplexity is, the better the performance of the model is, therefore, the K corresponding to the lowest perplexity or the inflection point will inflect the best number of subjects.

2.2 Hot Topics Identifying

In this section, referring to the index setting methods for detection of existing literature emerging topic, and considering the characteristics of LDA model outputs, combined with the external form and content characteristics of Chinese and international journal literature, a dual discriminant index of novelty index (NI) and strength index (SI) is constructed (Fan & Ma, 2014).

2.2.1 Novelty Index

NI (Novelty index) reveals the temporal dimension of the topic. Novelty judge the novelty degree of the subject based on the age of the subject, The calculation formula is shown in Formula 3:

$$NI_t=\frac{1}{y-t(i)+1}(0<NI<1) \quad (3)$$

Where y represents the current year and t (i) represents the occurrence time of the topic i ($t \leq y$), so the novelty of topic decreases with years. The slope of the index decreases with a concave curve, in line with the law of literature aging, that is, the literature in the emerging stage has the fastest aging speed and the highest elimination rate. With the development of time, the valuable literature is usually retained, and the elimination rate decreases, so the aging speed is gradually slowed down, and the slope is also gradually reduced.

2.2.2 Support Index

SI (support index) reveals the intensity dimension characteristics of topics and reflects the attention degree of topics. By the document topic matrix derived from LDA model, the probability distribution of topics in documents can be obtained, and a set of supporting documents related to topics can be obtained. In this paper, the threshold of decision probability of supporting documents is set to 10% (Mimno et al., 2006). If the composition

probability of a topic in a document is higher than or equal to 10%, the document is regarded as a supporting document of the topic. In the topic model, a document represents a document, so SI of topics in a same time slice is the proportion of the number of supporting documents of the topic in the total number of documents, which can also be named as topic strength, as shown in formula 4.

$$SI_y^d = \frac{Sum_d(d)}{Sum_d(y)} \quad (4)$$

Where y represents the current year, the denominator represents the total number of documents in the current year, and the numerator represents the number of documents supporting d in the current year. The higher the SI, the larger the number of supporting documents and the stronger the topic strength.

2.2.3 Defining Hot Topics

Considering that the novelty of each topic changes dynamically in different time slices, topic models are constructed for each time slices, and the dynamic evolution process of topics strength can be shown. In order to intuitively reflect the process, this paper refer to the existing method of mapping topic state in two-dimensional space, and combine with novelty and support index, setting the index threshold as the coordinate origin according to the actual situation, and draws the topic life cycle coordinate map (Liu et al., 2018). Thus, the topic evolution process can be divided into four periodic stages: potential stage, emerging stage, hot stage and declining stage. Each topic will go through these four continuous evolutionary processes, as shown in Figure 1.

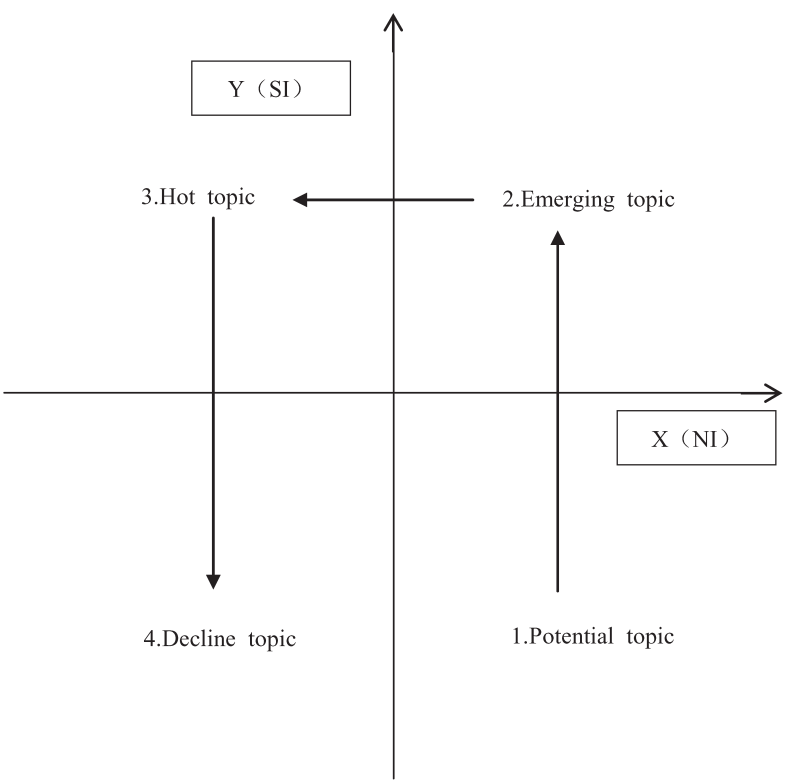

Figure 1 The Topic Life Cycle Coordinate Map

As shown in Figure 1, each quadrant represents a stage, in which each topic will be in the corresponding stage:

(1) Potential stage (the fourth quadrant): in this quadrant, topics are usually novel, but the SI is low, and the rate of supporting documents grow slowly, which indicates that topics have just appeared, but there are few related research, and they are in the embryonic stage.

(2) Emerging stage (the first quadrant): the growth rate of documents is gradually rising. In this quadrant, the NI and SI of topics are comparatively higher (less than hot topics), indicating that the related research of topics is gradually increasing and in a period of rapid development.

(3) Hot stage (the second quadrant): in this quadrant, the NI is low, but the SI is high, and the document growth rate is relatively stable, which indicates that the topic has been fully developed and the research heat continues to grow, and has entered the mature stage.

(4) Decline stage (the third quadrant): in this quadrant, the topic NI is low, and the SI is also very low, which indicates that the topic is older, the research enthusiasm is declining. The supporting documents of this kind of topics are gradually decreasing, and the direction of evolution is declining or changing into new research topics.

Therefore, the topics in the second quadrant can be defined as hot topics.

2.3 Topic Evolution Path Forming

Topics are extracted by LDA model, and the topics and keywords in the adjacent time slices turn out to have certain similarities and differences. By calculating the topic similarity of different time slices and setting a certain threshold, the topics with high correlation can be determined, and then the evolution relationship between them can be determined, so as to form the evolution path of this kind of topics.

In this paper, cosine similarity is used to measure the similarity between hot topics in different time slices, so as to determine the evolution relationship and form path. Cosine similarity is measured by cosine value of the angle between two vectors in vector space to compare the similarity of two individuals represented by vectorization. In the two-dimensional vector space, suppose two two-dimensional vectors: a vector is (x1, y1), b vector is (x2, y2), then the cosine theorem can be expressed in the following form:

$$cos(\theta) = \frac{a.b}{||a|| \times ||b||} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}} \tag{5}$$

In the same way, the vector is expanded from two dimensions to n-dimensions, and a and b are assumed to be n-dimensional vectors:

$$cos(\theta) = \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}} \tag{6}$$

The closer the cosine value is to 1, the more similar the two vectors are. If the value equals 1, the vectors are equal. Therefore, when it is applied to the topic in the form of word set, we can judge whether the topic is related or not by vectorizing the topic features, constructing a dictionary and establishing a vector space, and calculating and comparing the cosine value in pairs.

2.4 Topic Evolution Visualization

In order to show the updating, evolution and development of topic content, as well as the

trend of topic intensity changing with time, the visualization of topic evolution in this paper is based on LDA model and similarity, which are respectively displayed from two aspects of content and intensity.

2.4.1 Visualization of Topic Content Evolution

The visualization of evolution path of topic content is realized by Sankey diagram, also known as Sankey energy split diagram, which is originated from the "energy efficiency diagram of steam engine" in 1898 (Liu et al., 2018). where the element block represents the object, and the line represents the flow direction and connection of the energy generated by the object. Using this feature, we can intuitively show the updating and increasing of the topic content over time. The Sankey diagram of topic content evolution will be drawn by visualization software "InfoCaptor".

2.4.2 Visualization of Topic Idensity Evolution

The evolution of topic strength (support) is based on the number of supporting documents, which is shown by a line chart. In order to show the trend of topic intensity changing with time, count the number of supporting documents for each path in each time slice, and generate a line chart of topic intensity trend with abscissa of time and ordinate of number of documents.

3 Results

3.1 Data Sources

The field is limited to Digital Library. The data collection period start from the emergence of related research and application in the field of Digital Library. Starting from the year when more than two literatures appeared in the field, the Chinese starting from 1994, and the English start from 1991, both of which are up to 2020.Data is retrieved on December 15, 2020.

The Chinese literature resource come from China National Knowledge Infrastructure (CNKI), which is limited to SCI journals, EI source journals, core journals and CSSCI / CSCD, and the type of literature is set as journals. Retrieval method is professional search, set the search formula as "SU=Digital Library", where "SU" means Chinese subject. Download the full record of literature information and export it in Excel format in batch, and screen out the duplicate and incomplete literature records. Finally, there are 13,704 full records of literature.

English literature comes from WOS (web of Science), which is recognized as an authoritative literature index tool in the world. It contains rich and high-quality literature in many research fields. In this paper, professional search is used, and the search formula is "TS= ('Digital Library 'or' e-library ')", where "TS" stands for English subject, the type of literature is "article", and the language is limited to "English", including SCI and SSCI citation index. Download the full record of literature information and export it in Excel format, screen out duplicate and incomplete literature records, and finally there are 7,706 full records of literature.

In this paper, time slice segmentation is based on timeline. In view of the fact that the number of journals in the selected fields has increased greatly in the past ten years, but the number of papers published in the early stage is very low, in order to balance the number of literature records in each time slice, the records before 2000 are summarized into the same time slice, and the next 20 years are divided into 11 time slices with 2 years as a fixed length time slice.

3.2 Text Preprocessing

3.2.1 Corpus Sources

Selecting title, keywords, abstracts and other information from the exported literature information (Zeng et al., 2014). Title, abstract and keywords are selected as the corpus sources of the model. Due to the authoritative and complete information provided by the existing keywords and extended keywords in English literature records, in order to avoid the fragmentation of professional terms and semantic loss caused by word segmentation and stem extraction, the above two (keywords and extended keywords) are directly combined and used as the source of English corpus.

3.2.2 Word Segmentation

Chinese document is separately created by time slice. using Chinese word segmentation packet "Jieba" in precise mode to segment the text of each document. Before word segmenting, the keywords in the existing literature will be collected to form a user-defined dictionary in the field of Digital Library, so as to retain the professional vocabulary in the process of word segmentation. screen out function words and meaningless symbols. In order to unify the Chinese and English professional expression and abbreviation, a thesaurus will be set to merge synonyms. Due to the high degree of standardization and unification of substantive words in journal literature, it is difficult to define whether there is a close relationship between professional words. Therefore, it is not necessary to merge synonyms.
Keywords are directly used in English literature information without further word segmentation.

3.2.3 Dictionary Construction

The words usually possess strong semantic features after word segmentation. In order to convert it into numeric information and apply it to the topic model, this paper uses the method of bag of words model to build a dictionary. At first, all the words in the document are collected and duplication eliminated, and then each word be assigned an index (serial number, eigenvalue). Because LDA model is a statistical model based on word frequency, the eigenvalue will be set as word frequency.

3.2.4 Feature Representation

Using the dictionary constructed in the previous step, each word in documents will be represented by corresponding eigenvalues. Each line represents a literature information, and the words in each line will be represented by indexes, then all the literature information in documents will be transformed into an acceptable input format for LDA modeling.

3.3 LDA Modeling

For the document information generated in the previous step, the open source packet "Gensim" is used to build the topic model and estimate the parameters. Super parameters are selected as α =0.05、β =0.02 by default.

The optimal topic number k is determined by solving the perplexity of 11 time slices of Chinese (cn) and international (en) document according to formula 2, obtaining the average value and forming a line chart, as shown in Figure 2.

As can be seen from Figure 2, with the increase of the number of topics, the perplexity of the model decreases obviously in the early stage, which indicates that the model has good performance and can effectively distinguish topics. the K value (Optimal number of topics) corresponding to the lowest perplexity should be taken. However, in the case of small corpus, more topics may lead to over-fitting. Therefore, the K value corresponding to the

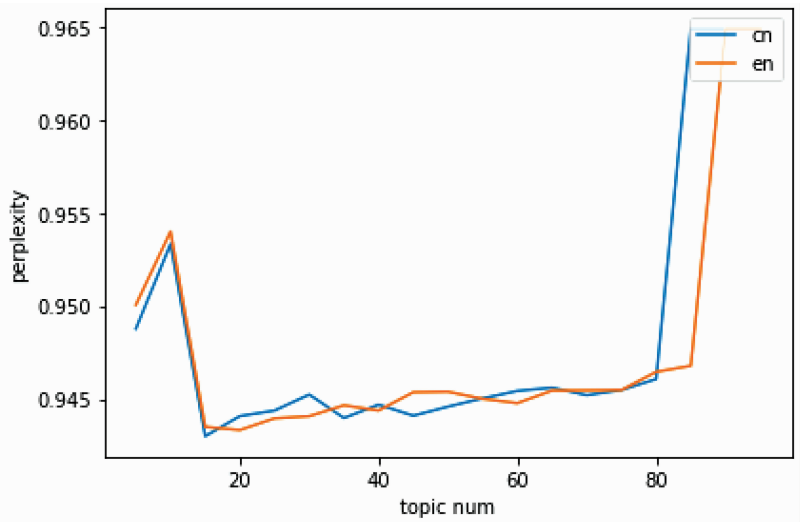

Figure 2 Perplexity

first inflection point will be taken which is 18 on average at home and abroad.

After the parameters of LDA model are determined, the Gibbs sampling algorithm is used to solve the distribution of the potential probabilities and for each time slice. Then two types of documents are generated for each time slice.

The first one is document-topic distribution, and the document is named "Doc_topics.csv" (exported in tabular format), including multiple topics contained in each document and the probability of each topic, which is mainly used for topic strength calculation. The second one is topic-word distribution, and the document is named "topic_words.csv". The number of keywords under each topic is limited to top-25, considering that the probability of keywords after 25 is too low.

3.4 Hot Topic Identification

Based on the topic information obtained by LDA model, the NI and SI of topics will be calculated, and the threshold will be determined according to the situation. The life cycle coordinate diagram of topics will be drawn, so the evolution stage of each topic can be determined, so as to identify hot topics.

3.4.1 Topic Lifecycle Thresholds

The classification of novelty refers to the "Pareto principle". There are 11 time slices, and 20% of the years account for about two time slices. Therefore, the NI of topics generated in 2015, which is five years before the statistical year, is used as the threshold to determine the temporal characteristics of topics. According to formula 3, the threshold value of NI come out to be 200 (the index value is expanded by 1000 times year-on-year for the convenience of comparison).

The threshold value of SI indicates the strength characteristics of topics, which requires solving SI of all the topic by formula 4, and taking the arithmetic average of them. After calculation, the threshold value of Chinese index come out to be 55.5, and that of international index is 54.4.

3.4.2 Hot Topic Recognition Based on Topic Lifecycle Evolution

Lifecycle feature of Chinese topics (CN) and international topics (EN) can be visually shown

in a coordinate axis, where the NI is taken as the x-axis, and the SI is taken as the y-axis. Due to the different thresholds of the SI at home and abroad, the left axis value of the y-axis represents the Chinese support value, and the right axis value represents the international support value. Set the threshold of NI and SI as the coordinate origin, determine the coordinate axis and the position of each topic on the coordinate axis, and draw the topic life cycle stage diagram, as shown in Figure 3.

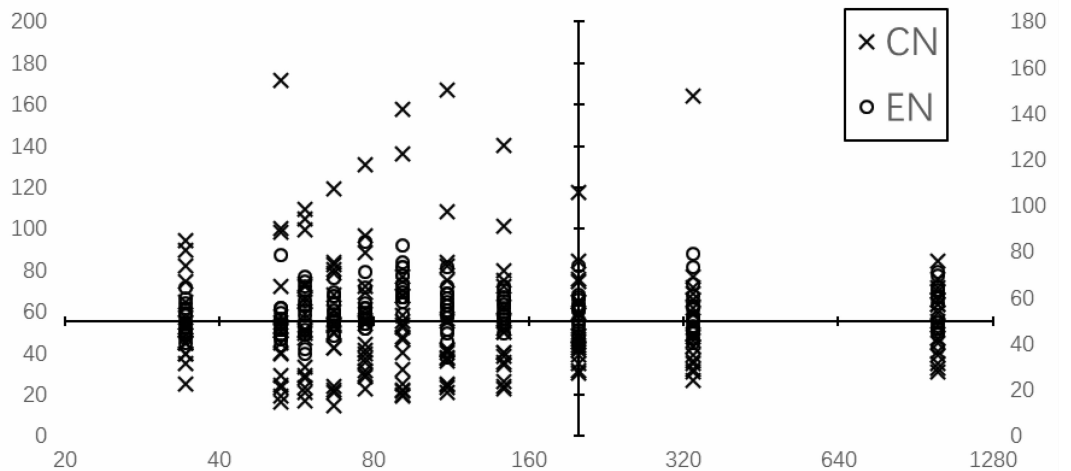

Figure 3 Topic Life Cycle Stage Diagram

The distribution of research topics in the field of Digital Library at home and abroad can reflect the structural evolution characteristics of the research, which is basically in line with the periodic distribution law of the topic evolution cycle: in the potential stage and emerging stage, the SI of topic is higher, but the topic with the highest SI is still in the hot stage. On the other hand, the NI shows a decrease in variance with the development of time.

In Figure 3, the SI distribution of EN is relatively closer to the x-axis, and the overall variance is less than that of CN, indicating that the research prevalence of topics is evenly distributed, and the topic structure equilibrium is better than that of CN; there are more Chinese potential topics, but less emerging and hot topics, indicating that the Chinese gained more momentum on innovation. Though there are many research fields involved, the scope is wide but not deep. The phased distribution of topic life cycle is shown in Table 1 below.

Table 1 Periodical Distribution of Topic Life Cycle

	potential topic	Emerging topic	Hot topic	Decline topic
Number of CN	91	53	23	31
Number of EN	77	67	24	30

Example of China and international hot topics are listed as follows: Table 2 and Table 3.

Table 2 Examples of Chinese Hot Topics

topic1	topic2	topic3	topic4	topic5
The construction of Digital Library	cloud computing	copyright	Collection resources	big data
College Library Working Committee	Cloud computing environment	intellectual property right	service	system
Sociology	Resource allocation	Intellectual property protection	Knowledge visualization	Virtual Reality Technology
Economics	user	Copyright	Collection of Digital Library	colleges and universities
representative	service	Copyright issues	improvement	information service
proposal	Mobile Digital Library	protect	knowledge service	field
China Digital Library Project	Personalized recommendation service	information technology	knowledge organization	Cloud model
information resources	Reasonable allocation	information resources	Open access	research hotspots
Beijing	Collaborative filtering	Storage technology	Visualization technology	key word
books	Personalized service	focus	readers	evaluation

From Table 2, we can find that topic1 is about the social background of Digital Library construction; Topic2 is the personalized service based on cloud computing; Topic3 is mainly about copyright related legal issues; topic4 is the knowledge service of collection resources; topic5 is the scientific research information service of big data realization.

Table 3 Examples of International Hot Topics

topic1	topic2	topic3	topic4	topic5
databases	user acceptance	digital libraries	literacy	gene
image and video databases	user evaluation	accuracy	cocitation	cell library
classification	performance	security	research work	rna recognition
text mining	networks	tracking	e–textbooks	gene–expression
segmentation	complexity	networks	information retrieval	journals
selection	personalization	errors	academic libraries	serial analysis
digital video	internet	attributes	reference services	records management
model	user studies	recognition	information objects	cancer
support vector machine	students	reliability	digital objects	knowledge
Neuron network	system	privacy–preserving	intellectual property	trends

Similarly, from Table 3: topic1 is about database and storage and related algorithms; Topic2 is network-based information services and users; Topic3 is information security related; topic4 is education and scientific research; topic5 is the application of Digital Library in biology.

3.5 Dimension of Topic Content and Identification of Topic Evolution Path

3.5.1 Dimension of Topic Content

In view of the technical characteristics in the field of Digital Library, the topic dependent

keywords can be separated into two dimensions of technology and application. The effect of keyword clustering from the technology dimension is better, but the application objects of keywords from application dimension are often too micro and scattered, and they often involve many cross-application fields, so it is difficult to construct evolution path by mixing technology and application keywords. Therefore, this paper analyzes the evolution of topic content from two dimensions: technology and application. The technology dimension [T] involves Digital Library algorithm, method, technology and theory; the application dimension [A] mainly refers to the application of Digital Library technology.

In the division of topic content dimensions, first of all, the keywords should be filtered and remove the noisy words (words with weak real meaning expression or fuzzy classification), and then the two dimensions should be manually annotated. The examples are shown in Table 4.

Table 4 Annotation Examples

topic4	copy-right [A]	intellectual property [A]	intellectual property protection [A]	Digital signature [A]	copyright issues [A]	privacy [A]	information technology [T]	information resources [T]	storage technology [T]	focus [A]

The topic content of each time slice is divided into two tables according to this standard, which helps to respectively identify technology and application evolution path.

3.5.2 Identification Topic Evolution Path

In order to construct the evolution path of topic, there should be a title of each topic path, which is called the main path identifier. The keywords under hot topics are arranged in order of probability. The keywords with higher probability in the front row usually represent larger concept category, reflecting the main research direction of the field; while the keywords with lower probability in the back row usually represent smaller concept category, studying specific technology, sub-theory and application object, which can represent the characteristics of the topic more specific. Therefore, the main path of technology dimension is identified as high probability keywords in each time slice.

In view of the keywords of application dimension with high-ranking probability are not general and suitable for classification, Chinese keywords refer to the Chinese Library Classification to identify and classify the main path, while English keywords refer to the research direction of web of science to identify and classify the main path.

The evolution relationship of topic content is obtained by calculating the similarity of topics in adjacent time slices through formula 6. According to the existing relevant research, 0.3 can be determined as the similarity threshold (Liu et al., 2016). Hot topics with similarity higher than 0.3 of adjacent time slices are judged to have evolutionary relationship, and should be classified into the same topic path. The content of the same topic evolution path is constructed by keywords under the same topic path in adjacent time slices.

4 Discussions

4.1 A comparative analysis of the evolution of Chinese and international hot topics in technology dimension

4.1.1 Visualization of Chinese hot topic evolution in technology dimension

The topic content and evolution relationship data are imported into "Info Captor" software

to generate the topic content evolution Sankey map. The bottom horizontal axis represents the time slice, the first column on the left is the main path identification, the element block in the following column represents all the emerging topic content (keywords) under different time slice, and the gray line connects the keywords belonging to the same topic path. In order to highlight the evolution and update of topic content, keywords are only retained in the time slice when they first appear. Therefore, with the passage of time, the number of keywords in each topic path and each time slice decreases.

The first column of the topic is the topic identifier, and high probability keywords under hot topics are selected. the horizontally connected content is the new content of the topic in different time slices. According to the results of LDA topic extraction from Chinese corpus, the main paths topic identifier can be identified as keywords with probability greater than 0.1, and eight paths are identified, as shown in Figure 4.

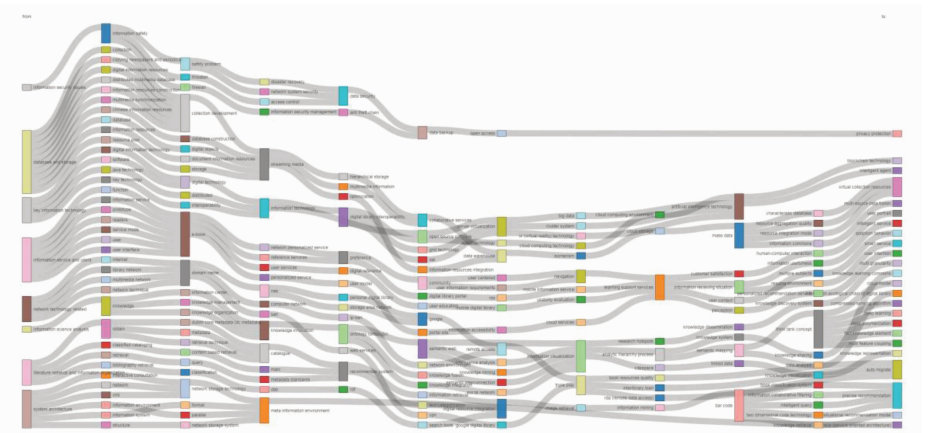

Figure 4 Visualization of Chinese Hot Topic Evolution in Technology Dimension

The total number of supporting documents of each topic is regarded as topic strength, and it can be plotted on the coordinate axis with the horizontal axis of time and the vertical axis of strength to form the evolution diagram of topic path strength, as shown in Figure 5. It can be found that the topic path intensity of Chinese technology research presents a

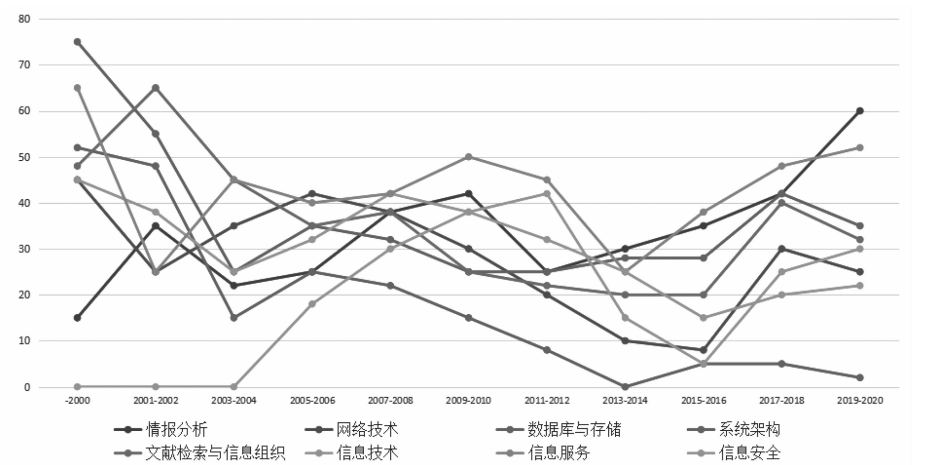

Figure 5 Evolution Diagram of Chinese Topic Path Strength in Technology Dimension

"U-shaped" trend as a whole, which decreases at first, and then bottoms out and rebounds in 2013-2014. A technical framework for new disciplines is built at first, and then technical shortcomings and innovation points comes out in the process of practice. In the case of new demands in the new era, the research focus turns back to the improvement and innovation of technology development. For example, in the early stage, China paid more attention to the basic technology and theory of database and system architecture, and now it pays more attention to information service, intelligence analysis and other aspects, focusing on the development from technology on back-end to service on front-end, and the overall trend of topic intensity evolution path is relatively consistent.

4.1.2 Visualization of International Hot Topic Evolution in Technology Dimension

According to the results of LDA topic extraction from English corpus, the topic words with probability greater than 0.04 can be selected as the main path identifier, and 9 main paths can be identified, as shown in Figure 6.

Figure 6 Visualization of International Hot Topic Evolution in Technology Dimension

Similarly, when international topic paths are combined with the intensity of each stage, a broken line chart of the evolution trend of topic intensity can be drawn, as shown in Figure 7.

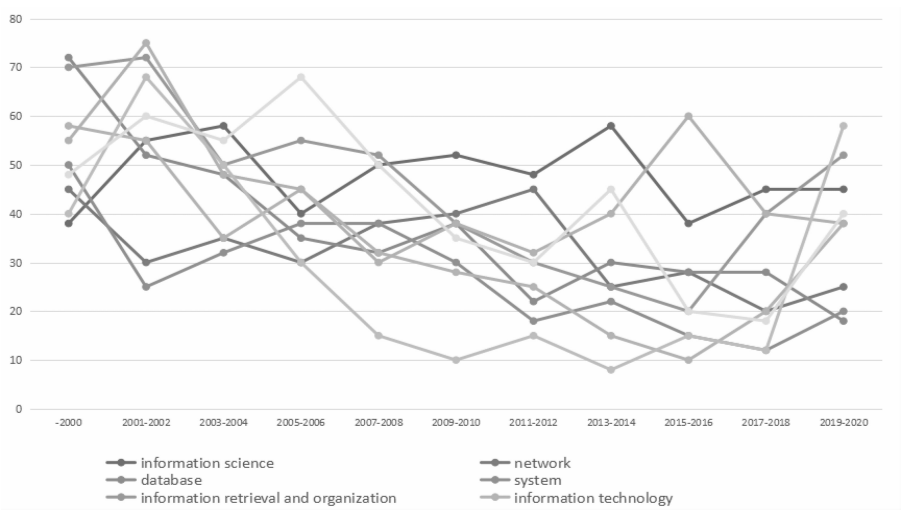

Figure 7 Evolution Diagram of International Topic Path Strength in Technology Dimension

As can be seen from Figure 7, from the overall trend, the change trend of topic intensity of international technology dimension is similar to that of Chinese, but less obvious. On the one hand, it is because trend of each topic intensity path differs a lot, such as "system" bottomed out far earlier than "information". On the other hand, the trend of some topics is stable, such as "information science", whose prevalence has been kept at a high level, while "database" continues to decline. It shows that the overall prevalence is higher than that in China, and the trend is more stable than that in China.

4.1.3 Comprehensive comparative analysis

On the whole, the research of technology dimension shows a downward trend and recently transfers to application research, but the trend of the overall intensity of international topics is more stable than that of Chinese topics, and keeps getting more attention. It shows that international research pays more attention to technology dimension in this field, and the technical foundation may be more solid.

Chinese and international structure and research direction of the Digital Library field are roughly similar. Combined with the internal structure of the field, there are the same eight evolution paths in China and internationally, while there is one more "algorithms" path abroad. Therefore, the following will compare the content and intensity of topic path from China and aboard in technology dimensions, and reveal similarities and differences in topic evolution process, so as to explores the causes.

"Information science" is to extract valuable information, through a series of processing processes of information collection, sorting, identification, evaluation, analysis and synthesis. The purpose of Digital Library is to facilitate users to obtain the required information more effectively. Therefore, information analysis technology for information processing is the guarantee to improve the quality and efficiency of operation. Libraries pay attention to the mining and providing of knowledge, so in the early stage, they pay more attention to knowledge organization, management and innovation. With the development of multi-source data, it also promotes the integration of knowledge. Combined with the generation of massive data, the concept of semantic web and think tank has been put forward, and deep learning, knowledge discovery and other technologies have begun to develop in depth. The concepts of information visualization and data association were put forward in foreign countries at the beginning of 21th century. The concepts of ontology, semantic web and hierarchical model were put forward about two years earlier than those in China. Nowadays, knowledge mapping and network analysis are new research hotspots.

"Network technology" paid more attention to the configuration of basic network system in the early stage, and the research on network technology abroad was more detailed and specific. With the development of user demand from local to international, the demand for network interconnection in different regions and types became larger, and the research on data interconnection between libraries and international countries gradually increased. In the later stage, the combination of network technology research and Digital Library got closer.

"Database", the biggest difference between Digital Library and traditional library is the digitization and systematization of a large number of existing text data. Therefore, information database technology must be constantly updated with the change of information form and quantity. At the beginning of the 21st century, the concept of Digital Library has just been popularized. more attention is paid to how to digitalize the existing collection resources. Then, there come the diversification of information forms (such as multimedia information), and development of big data and requirement for data resource integration (massive data, heterogeneous data, multi-source data), the emergence of new

information technology (such as cloud storage), they all put forward new requirements for Digital Library database construction and storage related technologies. At the beginning of the century, the types of data involved in foreign research are more diverse and complex, and tend to focus on image processing technology.

As another bottom architecture of Digital Library, "System" has been very prevalent before 2005. Later studies tend to be user-oriented and evaluation.

As an important methodology of traditional library, "information retrieval and organization" has changed greatly with the change of information carrier. This branch of technology field has the strongest topic intensity. The basic indexing, classification and retrieval technologies are comprehensively involved at the beginning of the century, but the further focus on metadata, feature extraction, semantic retrieval appeared earlier abroad, and the further development of these technologies (such as multi-dimensional retrieval, interactive retrieval, image index, etc.) is more in-depth and innovative abroad.

"Information technology" mainly refers to all kinds of computer technology used to manage and process information. With the continuous iterative update of hardware and communication technology, the related technology also got a high update rate. In the early stage, both China and foreign countries paid attention to the research of software and hardware, and foreign countries also laid the research foundation of distributed, parallel computing and virtual reality technology earlier. After 2010, both China and foreign countries began to pay attention to the development of cloud computing and artificial intelligence technology. Under the branch of information technology, foreign countries also include a branch of "Algorithms", which is a branch ahead of China. At the beginning of the century, it includes algorithms such as sequence, probability, neural network, etc. the corresponding mathematical knowledge research is also more in-depth, which indicates that Chinese technology field lacks the corresponding theoretical research.

"Information service", as a user-oriented system, has become a relatively large research branch in the people-oriented technological environment. In the early stage, more attention was paid to the realization of the underlying technology, and less attention was paid to the "people-oriented" technology optimization. Therefore, at the beginning of the century, the research only stayed in the basic concepts of information services such as users, needs, and interfaces. With the development of technology and the diversification of users' needs in the later stage, various personalized and humanized technology optimization were promoted, which has become more advanced in China in recent two years. In combination with big data and algorithm updating, it is advocated to provide intelligent services based on human's feelings.

"Information security", the digitization of resources and users' information, is bound to be accompanied by information risk. This branch in foreign countries has attracted more attention about four years earlier than that in China, which is related to the higher popularization rate of information technology abroad in the early days, so that users have a better awareness of information security. But in the later period, it has also been fully developed in China.

4.2 A comparative analysis of the evolution of Chinese and international hot topics in application dimension

4.2.1 Visualization of Chinese Hot Topic Evolution in application dimension

According to the main practice areas involved in the process of Digital Library implementation, results come out to be eight paths. Among them, the emergence period of

individual paths is relatively late, such as the large-scale application of medical science, library and information science, and education, which appeared in 2003-2004, while remote sensing surveying and mapping, and traditional Chinese medicine appeared in 2009-2010, as shown in Figure 8.

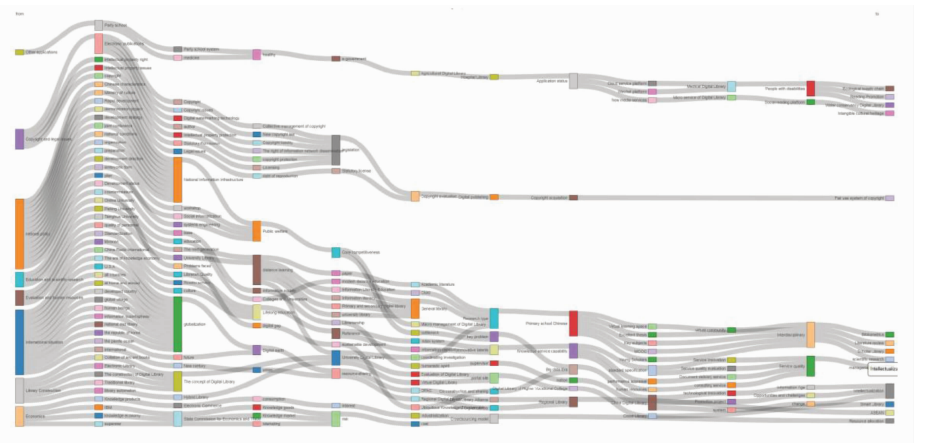

Figure 8 Visualization of Chinese Hot Topic Evolution in application dimension

Counting the supporting documents of each topic path to form a description of its strength, and generate a line chart of the evolution trend of topic strength of Chinese application dimension, as shown in Figure 9.

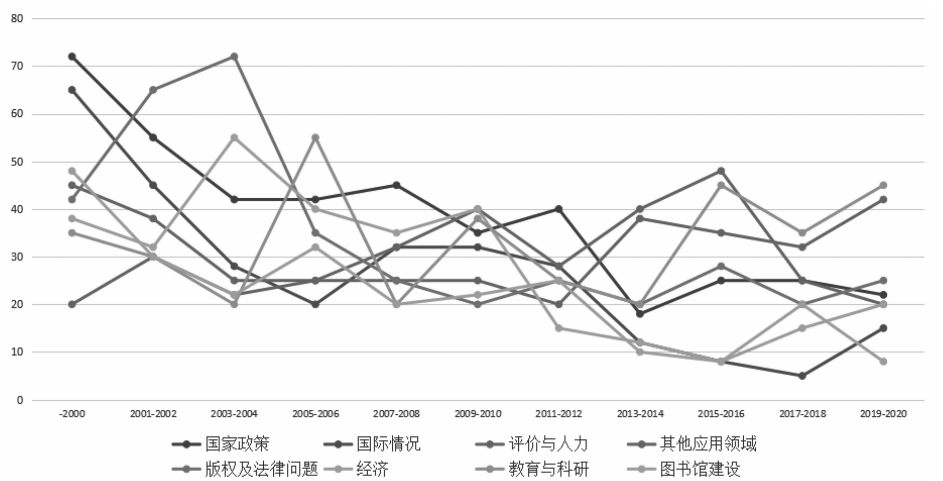

Figure 9 Evolution diagram of Chinese topic path strength in application Dimension

As can be seen from Figure 10, the distribution of application intensity in Chinese Digital Library field is relatively average in all application fields. In addition to the prominent distribution of copyright legal issues before 2003, the overall trend is fluctuating and declining. In the early stage, Digital Library field paid more attention to national policies, international situation, copyright legal issues and other hot spots, and with the development of time, it was applied more in education and other fields. According to the conclusion of

the previous section, the Digital Library field turns to pay more attention to the development of technology in the later stage, and the overall attention on application dimension is reduced.

4.2.2 Visualization of International Hot Topic Evolution in application dimension

main path identifier references category of Web of Science, combining with the main practice fields involved in the implementation process of Digital Library, 11 paths are identified as the main path, as shown in Figure 10.

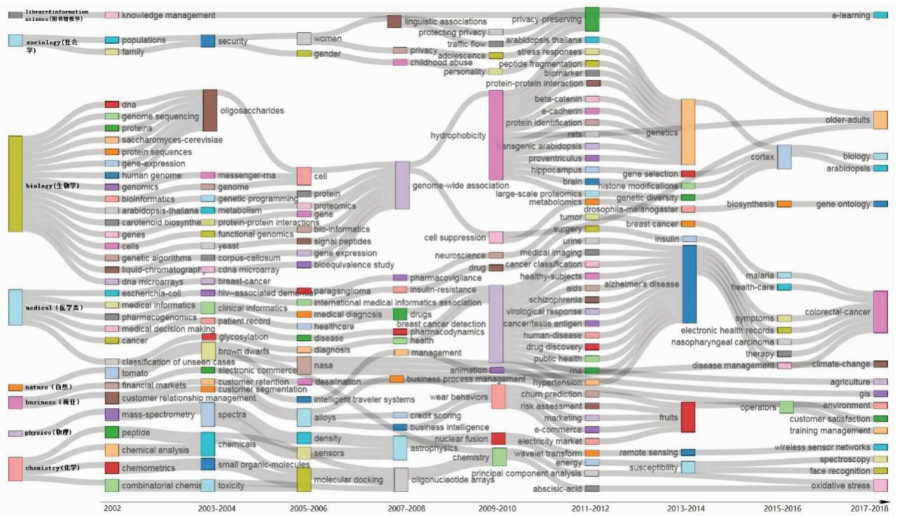

Figure 10 Visualization of International Hot Topic Evolution in application dimension

Counting the supporting documents of each topic path to form a description of its strength, and generate a line chart of the evolution trend of topic strength of international application dimension, as shown in Figure 11.

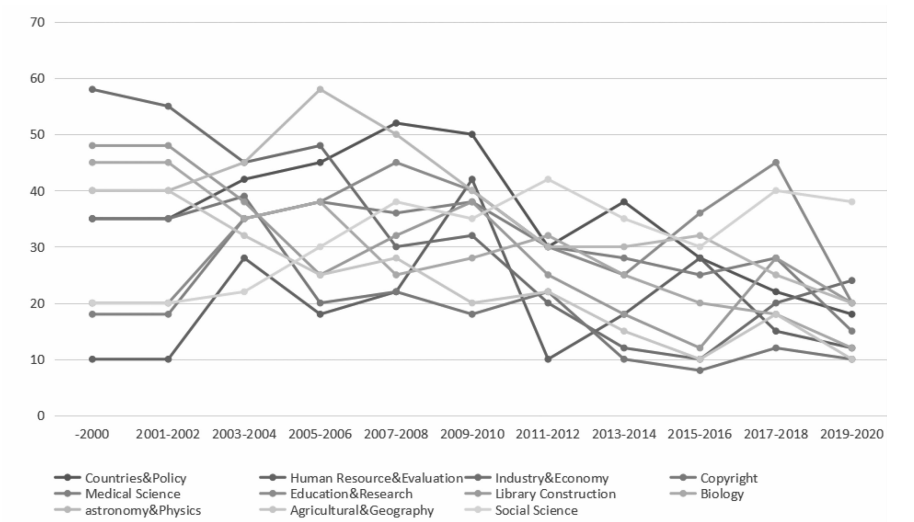

Figure 11 Evolution diagram of international topic path strength in application Dimension

As can be seen from Figure 11, the intensity distribution trend of international Digital Library applications is similar to that in China, they overall decreased in a fluctuation way. In the early years, it was widely used in the fields of biology, medicine and other natural sciences, but in recent years, it has turned to pay more attention to the application in the fields of education, humanities and social sciences.

4.2.3 Comprehensive comparative analysis

From the overall distribution, the application dimension of foreign digital libraries is more dispersed than that of Chinese ones, there are more categories and more keywords involved, which indicates that foreign digital libraries are involved in a wider range of application fields. From the details of the topic content, national policies, evaluation and manpower, copyright and legal issues, education and scientific research, economy and industry are involved both in China and abroad; but the application of biology, medicine, astrophysics, agricultural geography, social sciences and other majors is more in-depth on abroad. Most of the domestic application fields are social sciences and focus on the construction of Digital Library, while foreign countries are more emphasis on Natural Science and more in-depth application.

"Countries& policy" attaches great importance to the development strategy and planning of Digital Library at the beginning of the 21st century, and the Chinese development bears more political meaning. At the same time, from the international situation, we can see that China attaches great importance to international communication in the early stage, learning from the development experience of foreign countries. After 2010, China began to put forward nationalization, characteristics, reform and innovation of Digital Library, and put forward with humanism, intelligent development strategy earlier. On the other hand, the popularization of Digital Library in foreign countries is promoted from developed countries to developing countries. For the corresponding branch of "library construction", the concept of Smart Library was first proposed in China. The overall construction is closely combined with the field of education, focusing on integration and popularization, while in foreign countries, it is more combined with natural disciplines, focusing on detailed development.

"Human Resource& Evaluation" and "copyright", both in China and abroad have maintained a certain degree of attention, and the requirements have gradually improved over time. For example, in terms of human resources, from the beginning of basic personnel management, higher requirements have been put forward for technology, information literacy and service quality; in terms of copyright, the awareness of copyright has been continuously improved, and the improvement of relevant laws has been promoted.

As for "Education& research", both in China and abroad pay more attention to the role of Digital Library in education. In the early stage, it is mainly reflected in distance education and campus library construction. In the later stage, it plays an important role in scientific research, which can effectively promote the sharing of literature resources and scientific communication.

In terms of interdisciplinary and natural science, there are few applications in agriculture and medical treatment in China. While in foreign countries, there are far more applications in biology, medicine, astrophysics and agricultural geography, especially in biology and medical treatment. From the perspective of social science, in China, most of social application is based on the popularization of Digital Library and the media promotion, while foreign countries pay less attention to the promotion activity, and focus on specific social problems, art, culture, politics and other aspects, indicating that the application on social science will be

greatly affected by the political and cultural background.

5 Conclusions

Based on LDA topic model, this paper extracts the research topics from Chinese and international core journal papers in the field of digital library, detects hot topics based on the topic life cycle theory, and manually divides the topic content into two dimensions: technology and application, then calculates the similarity to build the evolution path in two, and also visualize evolution process of hot topic content and intensity. Finally analyzed them.

This paper reveals the differences and gaps of hot topic structure and content details in the field of digital library. On the one hand, the distribution of hot topics in foreign countries is more balanced, which indicates that the internal structure of field research is more balanced and stable. There are more potential topics in China, but less emerging and hot topics than in abroad, which indicates that there are many fields involved in China, but not deep.

On the other hand, in terms of research content structure, there is a "U-shaped" trend in technology dimension: in the early stage, they all focus on technical research, in the middle stage, there are more research in application dimension, and in the later stage, they turn back to focus on technical research. In technology dimension China can keep up with the international pace and popularize the basic technology and architecture in the early stage, but it is slightly backward in technological innovation. At the same time, there is a great lack of detailed and in-depth research on technology and theoretical basis research on Mathematics and algorithm. In application dimension, foreign countries have more extensive and in-depth application in the field of natural science, while China pays more attention to the construction of digital library, development direction of application dimension is greatly influenced by social culture and political background. The overall research direction in the field of digital library tends to integrate and deepen the development of technology and application, promote the integration of information resources, and put forward the concept of people-oriented and intelligent development, and carry out technological innovation.

The next step of research will be based on the conclusion that there is lagging effect on the topic content level. By quantitative analysis, the lagging degree of Chinese and international journals in the field of digital library will be clarified, the reasons for the lag will be explored, and relevant suggestions will be put forward.

References

Blei, D. M., & Lafferty, J. D.(2006). Dynamic topic models. *Machine Learning, Proceedings of the Twenty–Third International Conference(ICML 2006), Pittsburgh, Pennsylvania, USA, June 25–29,* 2006.

Blei, D. M., Ng, A., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3,* 993–1022.

Fan, Y., & Ma, J.(2014). Detection of emerging topics based on LDA and feature analysis of emerging topics. *Journal of the China Society for Scientific and Technical Information, 33* (7), 698–711.

Girvan, M., & Newman, M. E.(2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences, 99* (12), 7821–7826.

Liu, Z., Wang, X., & Bai, R.(2016). Research on Visualization Analysis Method of Discipline Topics Evolution from the Perspective of Multi–Dimensions: A case study of the big data in the field of library and information science in china. *Journal of Library Science in China, 42* (226), 67–83.

Liu, Z., Xu, H., Yue, L., & Fang, S. (2018). Research on lagging effect of topic diffusion evolution face to prediction of research front. *Journal of the China Society for Scientific and Technical Information*, (10), 979–988.

Kleinberg, J.(2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7* (4), 373–397.

Mimno, D., McCallum, A., & Mann, G. S.(2006, June). Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE–CS Joint Conference on Digital libraries (JCDL´06)* (pp. 65–74). IEEE.

Qiu, J. P., & Lv, H.(2013). The hot domain, research fronts and knowledge base of international library and information visual analysis of 17 journals´ knowledge map. *Document, Information & Knowledge, 3*, 4–15.

Wallace, M. L., Gingras, Y., & Duhon, R.(2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology, 60* (2), 240–246.

Wang, X., Bai, R., Wang, X., & Zhu, N.(2013). An automatic classification system of mass online academic literatures. *Library and Information Service, 57* (16), 117–122.

Wei, X., & Croft, W. B.(2006). LDA–based document models for ad–hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retriev*al (pp. 178–185).

Xu, S., Shi, Q., Qiao, X., Zhu, L., Jung, H., Lee, S., & Choi, S. P.(2014). Author–Topic over Time (AToT): A dynamic users´ interest model. In *Mobile, ubiquitous, and intelligent computing* (pp. 239–245). Springer, Berlin, Heidelberg.

Zeng, Li., Li, Z., Tan, Y.(2014). Analysis of topic evolution in scientific literature based on dynamic latent Dirichlet allocation. *Software (05)*, 102–107.

Zhang, Q., & Ma, F.(2007). On paradigm of research knowledge management: A bibliometric analysis. *Journal of Management Sciences in China, 6,* 65–75.