Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

Search Method of Web Traditional Forestry Culture Teaching and Education Resources Based on Fuzzy Set

Mingrui, Han[1*]
[1]Henan Polytechnic Institute, Nanyang, Henan, China
*Corresponding Author.

Abstract:

Internet contains a lot of traditional forestry culture teaching resources. However, due to the lack of effective retrieval tools, the majority of students have a poor understanding of its application. This paper presents a design scheme of traditional forestry culture teaching resources subject retrieval system based on Web. The system adopts the retrieval method of combining keywords with key modes and the search strategy based on system history records, which greatly improves the efficiency and accuracy of teaching information acquisition. This paper mainly focuses on the current mainstream PageRank algorithm, focusing on the calculation method of the algorithm and the influence of page link structure on PageRank value. This paper analyzes the effect of the algorithm in several models, such as independent website, including inbound link and outbound link, and puts forward its optimization strategy in traditional culture teaching resource retrieval. The experimental results show that this method can improve the efficiency and accuracy of web search for traditional culture teaching resources.

Keywords: *Traditional forestry culture, teaching resources, web search strategy, PageRank algorithm.*

I. INTRODUCTION

With the in-depth development of China's education information work and the rapid expansion of teaching resources, how to accurately and quickly search the required information from the massive teaching resource database has become an urgent problem to be solved in the process of teaching and learning [1]. Because the traditional search engines are mostly based on the simple matching technology of keywords, they lack the ability to understand the query conditions entered by users, return too much information, and the recall and precision are not

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

high, so that they have little significance to users. The teaching resources have their own characteristics, so it is an objective demand to study the special intelligent teaching resources search engine [2-3].

At present, specialized search for basic education resources in primary and secondary schools in China has not yet appeared [4]. Compared with the development of general document information retrieval, the development of education informatization reform and the construction of basic education resources in primary and secondary schools in China is relatively slow. When the wheel of history enters the 21st century, the rapid development of information technology accelerates the pace of education informatization, and the progress of Chinese information retrieval technology itself also brings good enlightenment to education informatization [5-6]. Ideal Information Technology Research Institute of Northeast Normal University is specialized in developing teaching support software and building teaching resources suitable for China's specific national conditions [7]. Relying on the experts and professors of Northeast Normal University, and serving the primary and secondary schools all over the country, they are exposed to a wide range of resources every day, and have higher requirements for the staff. They are required to understand the relevant professional knowledge as well as the relevant computer skills. Even so, it will inevitably lead to individual classification errors of resources [8-10]. The research and development of this search engine will undoubtedly play a great role in improving the work efficiency in the future.

To sum up, the research on search engine based on educational resources has important scientific research value and application value.

II. SEARCH ENGINE MODEL BASED ON WEB MINING

2.1 Research status of search engine based on Web Mining

The so-called Web personalization, in essence, is a kind of web service centered on user needs. Firstly, different web users access web resources through various ways. Secondly, the system learns the characteristics of users and creates a user access model. Finally, the system adjusts the service content according to the obtained knowledge to meet the personalized needs of different users. At present, there are several ways to store the user's interest information: one is to store the user's interest information on the search engine server; the other is to store the user's interest information on the user's machine. In this way, the user's interest information is stored on his own machine through cookie mechanism. When the user visits the corresponding search engine, the user's interest information and search keywords are sent to the search engine at the same time as the basis for the search engine to retrieve information. Third, the user's

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

interest information is not stored on the search engine server, nor on the user's machine, but stored on other servers. The cost is to increase the network load, and the query speed becomes relatively slow.

At present, the research on search engine personalization mainly focuses on the following aspects:

1. Major search engines have adopted various new technologies to provide users with more selection information, such as Yahoo! , ODP, Google adopts the method of providing users with document category hierarchy, so that users can select the category of interest first and then search for the next step, which greatly narrows the scope of query. Northern Light, WiseNut, Vivisimo are clustering displays that provide search results. Teoma not only provides clustering of results, but also provides optimization of query statements. The above method improves the search accuracy to a certain extent. However, the same queries submitted by different users all return the same results, and need more interaction from users.

2. Meta search engine and distributed information retrieval model

Through the selection of appropriate data sources and reasonable organization of search results to improve the search accuracy.

3. Information filtering technology and intelligent agent system

The main idea is to build explicit or implicit user personal information records, and use these records to recommend documents to users to make them more in line with the interests of users.

4. Personalized search technology

In the engines that use these technologies to optimize the query, some use user information to query; some do not learn user information, but use local relevant information to query; others need users to provide interest categories, select information sources and optimize according to the interest categories provided by users. A better way is that the search engine learns the user's search history, establishes user information profile and comprehensive profile, maps the request to two profiles when there is a new query request, and finally combines the query statement and category information to return the search results.

2.2 Model building of teaching resources search engine based on Web Mining

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

This research is based on the East Normal ideal teaching resources search project, mainly for the political subjects in the teaching resources database. At present, with the development of primary and secondary school information-based basic education and the increasing abundance of teaching resources, the backward Chinese search engine technology is becoming the bottleneck of autonomous learning, resource distribution and other needs. From the perspective of students, the use of search engine brings a lot of redundant information, which wastes a lot of their time; at the same time, due to the mass of information, it also adds a lot of inconvenience for software developers to find relevant information. In order to make full use of the existing teaching resources and meet the needs of information retrieval, it is urgent to build an intelligent search engine for basic education in primary and secondary schools. Here, we refer to the NENU ideal intelligent search engine system.

(1) System model

The system block diagram is shown in Figure 1: it is mainly composed of collector, controller, original database of teaching resources, indexer, searcher, user interface and other key parts.

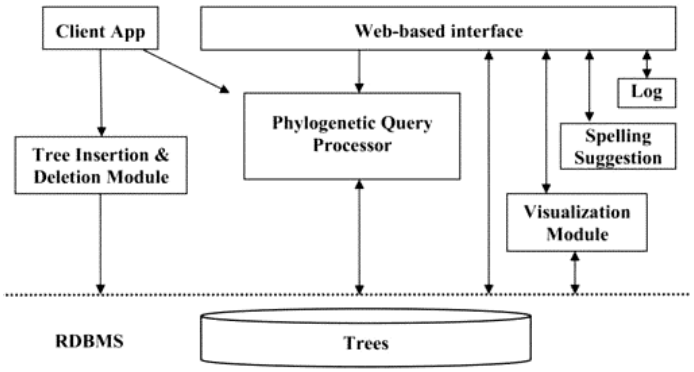

Fig 1: The architecture of ideal teaching resources search engine

(2) Collector

That is, the web spider, which searches for web pages through the link addresses of web pages, starts from a certain page (usually the home page) of the website, reads the content of the web page, finds other link addresses in the web page, and then searches for the next web page through these link addresses, so as to continue the cycle until all the web pages of the website

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

are captured. If the entire Internet as a website, then the network spider can use this principle to capture all the web pages on the Internet.

(3)Control clamor

It mainly solves the problems of comprehensive efficiency, quality and politeness. The so-called efficiency, here is how to use as few resources as possible (computer equipment, network bandwidth, time) to complete the scheduled amount of web page collection. Here we need to point out: even if we use a computer to collect web pages, we should also pay attention to the development and utilization of concurrency. And it's not that the more devices, the better. Under the arrangement of forming a cluster with several computers, they share the bandwidth of the export network. With the increase of the number of devices, the network bandwidth (or the bandwidth of some surrounding environment) will soon become a bottleneck. In addition, for the server, it may not be enough time to provide the required web pages. Therefore, do not let the crawler process started by the collector concentrate on a few websites. Focusing too much attention on several websites, or grabbing too many pages from one website in a short period of time, may also cause other serious consequences, that is, the so-called "Politeness" problem. It means that the normal crawling of web pages by web crawlers should not be too frequent to affect the normal access of website users. According to reports, Taobao (http: / / ww. Taobao. Corn) once caused the instability of Taobao server because the web spider of Yahoo search engine visited its webpage too frequently. The so-called quality problem refers to collecting Limited Web pages in a limited time, hoping that they will be "important" web pages as far as possible, or not missing those very important web pages. The importance judgment of web pages is also an important content of web structure mining. The commonly used algorithms are PageRank algorithm and hits algorithm. Generally speaking, the PageRank value of web pages closer to the home page is usually higher. In this way, it should be a better strategy to get as many homepages as possible first, and then to search from the homepages first.

(4)Retriever

The query service includes receiving the query phrase input by the user, searching, obtaining the corresponding matching result and displaying it to the user. At this point, we already have the index web page library and inverted file. What we need to do is to realize the intercommunication of index data and user query through the query agent. After information preprocessing, the data transferred to the service stage includes index web database and inverted file, which includes inverted table and index thesaurus. The query agent receives the query phrase input by the user, divides it, retrieves the document containing the query phrase from the index thesaurus and inverted file, and returns it to the user.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

The commonly used information retrieval models of retrievers are Boolean retrieval model, vector space retrieval model and probability retrieval model.

Considering the ideal reality of the East Normal University, the Boolean information retrieval model can not meet the requirements of fuzzy matching, and the probabilistic information retrieval model requires a large number of training data for parameter training. Therefore, it is decided to choose the vector space information retrieval model. The advantage of the vector space information retrieval model is that it simplifies the document content to the vector representation of the feature item and its weight, and simplifies the processing of the document content to the vector operation, which greatly reduces the complexity of the problem. The weight calculation can be done manually by rule method and automatically by statistics method, which is convenient to integrate the advantages of statistics and rules. And its query results can be sorted.

III. APPLICATION OF WEB STRUCTURE MINING IN TEACHING RESOURCES SEARCH ENGINE

In recent years, Google search engine has become the overlord of SEO industry and one of the most frequently used search engines in the world every day. This is not only because it provides an independent Cache system, dynamically generates summary information, and sets up a decentralized system (thousands of Linux clusters) for high-speed retrieval, but also because of the correctness of its search results, which is better than other search engines in ranking search results, which helps users find the required information as quickly as possible. PageRank technology contributed to google's great success. It was put forward by Sergey Brin and Lawrence Page, Ph.D. students from Stanford University in 1998 (both of whom are the founders of Google).

Firstly, PageRank algorithm is based on the following assumptions:

Let's assume that the network is a directed graph, G=(V, E), where V is the set of nodes (web pages) and e is the set of edges (edges from node I to node j exist if and only if there is a link from page I to page j).

The user randomly visits a webpage in the webpage collection at first, and then browses the webpage forward following the outward link of the webpage without backward browsing. All the clicking behaviors of the user are random and do not care about the specific content and theme of the webpage. The probability of browsing the next webpage is the PageRank value of

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

the browsed webpage.

If a web page is referenced many times, it may be very important. Although a web page has not been referenced many times, it may be important if it is referenced by an important web page: the importance of a web page is evenly transmitted to the web page it references. This important web page is called Authoritive web page. The basic idea of PageRank is to measure whether a page is important, or there are many links to it, or the page to it is important, or both. The initial definition is as follows:

$$PR(q) = \sum_{(p,q)\in E} \frac{PR(p)}{N_p} \qquad (1)$$

Where Np is the outgoing degree of node P.

When calculating PageRank, it is generally regarded as a process of finding the eigenvector of a matrix: M represents the transition matrix of G. if there is an edge from node j to node i, the value of elements Mi and j in the matrix is set to 1 / Nj, otherwise it is set to 0. In this way, the final result is: x = MX. Where x is the vector composed of PageRank of each page. From the composition of M, the maximum eigenvalue of matrix M is 1, and x is the corresponding eigenvector of 1. In this way, a simple iterative method can be used to solve the above equation.

If there are two mutually directed web pages a and b, which do not point to any other web pages, and there is another web page c, which points to one of a and b, such as a, then in the iterative calculation, the rank values of a and b are not distributed but continuously accumulated, as shown in Figure 2:

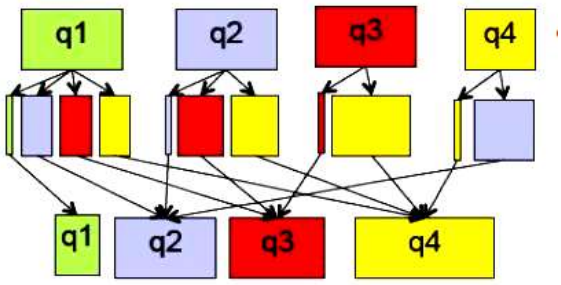

Fig 2: Non convergent PageRank iteration

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

To ensure the convergence of the above iterative process, M must satisfy two conditions: one is that the digraph G must be strongly connected; the other is that M must be acyclic. The latter can be guaranteed by the network structure, and the former can be guaranteed by adding a damping factor in the iterative process.

$$M' = cM + (1-c)\left[\frac{1}{N}\right]_{N\times N} \quad (2)$$

Using m' instead of m for operation is equivalent to adding two edges between every two nodes of g, which also solves the so-called Rank Sink problem. the iterative form at this time is as follows:

$$PR_{i+1} = cM \times PR_i + (1-c)\times\left[\frac{1}{N}\right]_{N\times 1} \quad (3)$$

Lawrence page proposed a random surfing model of user behavior to explain the above algorithm. They think that the probability of users clicking on links on the page is completely determined by the number of links on the page, which is also the reason for PR (Ti) / N (Ti) above. The probability of a page arriving by random surfing is the sum of the click probability of the links on other pages. Damping coefficient c is introduced because it is impossible for users to click on links indefinitely, and they often jump into another page randomly due to fatigue. C can be regarded as the probability that users click down infinitely, and (1-c) is the page level of the page itself. Observe the link structure of the web page as shown in Figure 2:

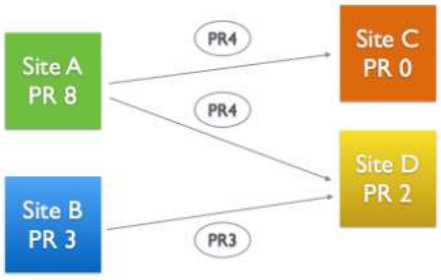

Fig 3: PageRank calculation diagram

IV. RESEARCH ON CLUSTERING TECHNOLOGY BASED ON WEB TEXT MINING

With the rapid development of the Internet, complex resources are gathered on the network,

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

including text, image, sound and other information. Of course, the most of them are text class data. According to the 19th statistical report on the development of China Internet issued by CNNIC on January 23, 2007, only 4109020 domain names have been reached in China, with 843000 websites and 4.47 billion national pages. "How to find the knowledge needed by many texts is an important problem to be solved at present, and it also makes text mining a hot topic in theoretical research. For our ideal Information Technology Research Institute, the main data source of teaching resource base, especially the teaching resource base of political subject, is also text. If there is such a method, when search engine climbs several pages, it is very beneficial to use text mining technology to classify these pages automatically, which is undoubtedly beneficial to reduce manual intervention and improve work efficiency.

The so-called text mining refers to the computer processing technology of extracting valuable information and knowledge from text data. It is a branch of data mining. Text data mining is an interdisciplinary subject, which is composed of machine learning, mathematical statistics, natural language processing and other disciplines. Text data mining is application driven. It is widely used in business intelligence, information retrieval and bioinformatics. For example, customer relationship management, web search and so on. In 1995, Ronen Feldman published "knowledge discovery in textual database" on KDD95, which is widely regarded as the first research literature of text mining, and also marks the rise of the field of text mining.

Our understanding of text data mining can be illustrated by Figure 4. This graph consists of three parts: the bottom layer is the basic field of text data mining, including machine learning, mathematical statistics, natural language processing; on this basis, it is the basic technology of text data mining. There are five categories, including text information extraction, text classification, text clustering, text data compression and text data processing; on the basis of basic technology, there are two main application fields, including information access and knowledge discovery; information access includes information retrieval, information browsing, information filtering and information reporting; knowledge discovery includes data analysis and data prediction.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

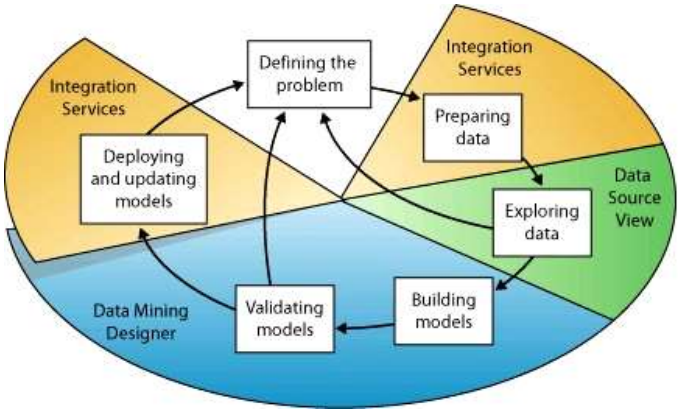

Fig 4: Text data mining

As can be seen from the above figure, text clustering is one of the important technologies in text data mining, which refers to classifying text according to its characteristics. That is to say, the given text set is divided into several subsets, which are called classes, so that the text within each class is similar, but the text between classes is not similar. The features of text are often different according to different applications. The similarity between texts is often determined by the application.

4.1 Overview of clustering

Clustering is to classify a group of individuals into several categories according to their similarity. Its purpose is to make the distance between individuals belonging to the same category as small as possible, and the distance between individuals of the same class as large as possible. Clustering methods include statistical method, machine learning method, neural network method and database oriented method.

In statistical methods, clustering is called clustering analysis, which is one of the three methods of multivariate data analysis (the other two are regression analysis and discriminant analysis). It mainly studies clustering based on geometric distance, such as Euclidean distance, Minkowski distance and so on. The traditional statistical cluster analysis methods include systematic clustering, decomposition, addition, dynamic clustering, ordered sample clustering, overlapping clustering and fuzzy clustering. Statistical clustering analysis is a kind of clustering method based on global comparison. It needs to examine all individuals to determine the classification. Therefore, it requires that all data must be given in advance, instead of adding new data objects dynamically. Clustering analysis method does not have linear computational complexity, so it is difficult to apply to the case of very large database.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

In machine learning, clustering is called unsupervised or unsupervised induction. Compared with classification learning, there is no class mark for clustering instances, which needs to be determined automatically by clustering learning algorithm, while there are class marks for clustering instances or data objects.

In neural network, there is a kind of unsupervised learning method: self-organizing neural network method, such as Kohonen self-organizing feature mapping network, competitive learning network and so on. In the field of data mining, neural network clustering method is mainly self-organizing feature mapping method, which can be used to cluster and segment the database.

4.2 Characteristics of Post search clustering algorithm

Many document clustering algorithms rely on offline and pre search clustering of the collected documents. However, because there are too many web pages and they are not always fixed, they can not provide offline clustering very well. When indexing, some search engines use a set of fixed topic marks to mark all documents, and then display the query results about this topic mark. Although this method is fast, this method of clustering before searching is local: these classes are not determined by the "local" pattern of the result set, but by the "global" pattern of all collected documents. Therefore, these classes are not important for the current result set. Another disadvantage of this method is that the pre-determined topic mark may not be suitable for the user's search target (on the other hand, the classes in the Post search clustering method are usually based on the features of the searched document set).

There are three key requirements of the searched document clustering system:

1. Relevance

This system should be able to generate classes, so that similar documents can be clustered (or in an information retrieval method, it states that the relevance of documents and user queries should be clustered).

2. Browsable summary

Users can know at a glance whether the content of a class meets their own needs. The system must provide concise and accurate descriptions of these classes.

3. Rapidity

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

An online system should have little or no significant delay for users. A patient user can filter the 100 documents listed. We hope that users can browse at least a large number of documents through clustering. Therefore, this clustering system should be able to cluster a large number of documents in a very short time. For a impatient user, a fast and scalable clustering algorithm is needed. Another important feature of this system is growth: in order to save time, it can process the documents just obtained from the network in real time.

V. CONCLUSION

The author mainly studies the structure mining and text mining in Web mining. This paper focuses on the PageRank algorithm, which is widely used in search engines at present, and focuses on the calculation method of the algorithm and the influence of web page link structure on PageRank value. Then, it analyzes the effect of the algorithm in several models such as independent website, including inbound link and outbound link, and puts forward the corresponding optimization strategy. Finally, by summing up the advantages and disadvantages of PageRank, an improved PageRank algorithm is proposed and verified for the topic drift phenomenon. But because the algorithm only analyzes the link structure of the network, there is inevitably the problem of topic drift, especially when there is more than one topic in the web page, the phenomenon of topic drift is more serious. Therefore, it is still a long way to go to improve the algorithm to solve the topic drift problem by using the text information of web pages effectively or studying the inherent characteristics of network link structure more deeply.

REFERENCES

[1] Huang Jianbin, Shao Yongzhen. the Way out of College English Teaching Reform. Foreign Language Circles, 1998 (04): 20-22
[2] Hu Wenzhong, Sun Youzhong. Highlighting Discipline Characteristics and Strengthening Humanistic Education -- on Current English Teaching Reform. Foreign Language Teaching and Research, 2006, 38 (005): 243-247.
[3] Liu Lude. The Enlightenment of Problem-based Learning on Teaching Reform. Education Research, 2002 (2): 73-77
[4] Liang Dingfang. My View on Foreign Language Teaching Reform. Foreign Language Teaching Theory and Practice, 2001 (1): 8-11
[5] Zheng Xinmin, Jiang Qunying. a Study on "teacher Belief" in College English Teaching Reform. Foreign Language Circles, 2005 (6): 16-22
[6] Liu Lude. the Enlightenment of Problem-based Learning on Teaching Reform. Education Research, 2002 (2): 73-77

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 118-130
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

[7] Li Guojie, Cheng Xueqi. Big data research: a major strategic field of future science and technology and economic and social development -- Research Status and scientific thinking of big data. Journal of Chinese Academy of Sciences, 2012, 27 (6): 647-657

[8] Zhou Yuanqing. the Construction of Excellent Course Materials is an Important Measure of Teaching Reform and Innovation. China Higher Education Research, 2003 (1): 12-12

[9] Li Zhiyi, Zhu Hong, Liu Zhijun. Guiding the Teaching Reform of Higher Engineering Education with the Concept of Achievement Oriented Education. Higher Engineering Education Research, 2014, 000 (002): 29-34

[10] Ye Lan. Let the Classroom Radiate Vitality -- on the Deepening of Teaching Reform in Primary and Secondary Schools. Education Research, 1997 (09): 3-8