International Journal of Applied Mathematics and Soft Computing

Volume 6, Issue 1, April, 2020

Publisher: HongKong New Century Cultural Publishing House
Address: Unit A1, 7/F, Cheuk Nang Plaza, 250 Hennessy Road, Wanchai, Hong Kong
Tel: 00852-28150191
Fax: 00852-25445670

Contents

Automatic Retrieval Model of Template Web Topic Information based on Vector Space

Qun Gao
Business School, Shandong University of Technology, Zibo, 255000, China

Abstract: In order to better realize the accurate retrieval of massive information in web pages, an optimization method of template web page theme information automatic retrieval model based on vector space is proposed. Combined with the principle of feature collection, the feature entry and exit retrieval of template web page theme information is analyzed, and the collected results are output and sorted to realize the automatic retrieval processing of template web page theme information. Finally, experiments prove that the template web topic information automatic retrieval model based on vector space has higher accuracy and fully meets the research requirements.

Keywords: Vector space; Template; The theme of the webpage; Automatic retrieval

1. Introduction

With the continuous development of network technology, the automatic retrieval technology of webpage subject information is also making continuous progress. The requirements for distributed retrieval systems under the network environment have become more stringent. Because the information distributed in different places can only be connected and retrieved through lines, the information processing is too loose and it is difficult to realize unified management [1]. However, the original web page theme information database information retrieval technology has some problems such as poor effect and incomplete functions in both information and audio. According to the above viewpoint, this paper optimizes the template web topic information automatic retrieval model based on vector space [2]. The structure of the distributed database of web page subject information is improved, and the web page subject information type database information is automatically retrieved by using the position operator and the limit operator. Furthermore, it integrates high-end technical factors. Finally, the experiment proves that the template-based automatic retrieval model of webpage subject information based on vector space has high reliability and practicability in terms of information transmission and access.

2. Template Web Page Topic Information Automatic Retrieval Model

2.1. Web page topic information feature collection

The attributes of the vector space templated web page subject information distributed database are closely related to the attributes of the network [3]. With the increase of network users, the mode, efficiency and activation time of data transmission will delay the network, thus affecting the accuracy and reliability of templated web page retrieval. The requirements of the vector space templated web page database for information retrieval are as follows.

(1) In the process of searching web page subject information, the most reasonable searching path should be selected, which involves the delay of the network and the parameters of the system, and the input and output in the local database are also related to the CPU [4].

(2) In the process of template web page information retrieval, concurrent control should be consistent with the network environment [5].

(3) The network communication environment should also be used when managing the catalogue of webpage subject information.

(4) When searching and distributing the theme information of web pages, the mutual transmission distance between each node of the database should be considered.

Web topic information database is a logical combination of nodes in the network environment [6]. It is inevitable to encounter a network structure of resource sharing in the process of information retrieval [7]. Any terminal in this state can share information of the resource database [8]. According to the distribution of data, some are system files, others are files in the form of database structure, and each node is united by different kinds of databases. The web page topic information database structure has the following attributes:

(1) Resources containing logic and physical are used on a dynamic basis.

(2) Physical and logical components are distributed and can interact through communication networks.
(3) The resources of database information are shared with users, and there is no need to specify the location for service [9].
(4) Using advanced management mechanism to unify and control various distributed data.
The retrieval steps of the web page topic information database information retrieval technology are shown in the figure:

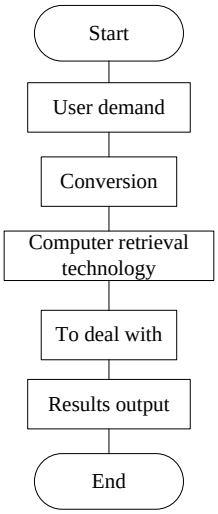

Figure 1. Retrieval steps of information retrieval technology

Information retrieval technology adopts the conversion technology of demand identification between computers and users, involving retrieval structure, program language, personalized demand [10]. Agent intelligence, information filtering, machine translation and other technologies. Information retrieval is based on the user's information questions, input questions into the computer, and then match the retrieved questions with the identification between the storage. Its core is mainly to meet the needs of the retrieval topic, and then calculate the identified retrieval expression [11]. However, the search expression can use the position operator and the limit operator to combine and match the key words of the question, thus determining the concept and position of the search key words, expressing the accurate content of the user's question, and ensuring the accuracy of the retrieval accuracy rate. The traditional dynamic data information resource retrieval system adopts a centralized serial type retrieval method, which is limited to the range of retrieval information and is difficult to realize the retrieval of dynamic data information resources in spatial databases [12]. As the sum of application-related dynamic data stored on computer physical storage media, spatial database is generally organized on the storage media in the form of a series of files with a specific structure. The design of dynamic data information resource retrieval system for spatial database should take network as the center and integrate all kinds of equipment to complete the retrieval work together [13]. Therefore, it is necessary to establish a data center and store a large amount of dynamic data information in the server so as to conveniently provide data services for the retrieval system. Its memory structure is shown in the figure.

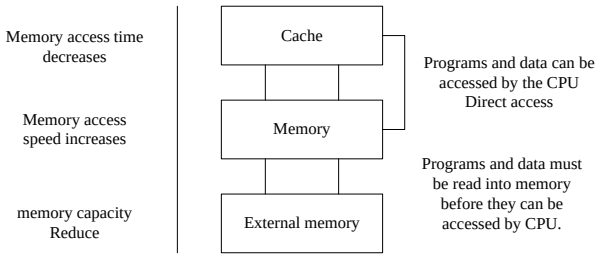

Figure 2. Dynamic data information collection

When the access time of the memory to data is reduced, the access speed needs to be increased, the access memory needs to be increased, and time is saved when dynamic data information is accessed, and the accessed data can be directly accessed by the CPU at this time; When the speed of memory accessing data is accelerated, relevant programs need to be used to read the data and convert it into the form of memory. At this time, the accessed data can be accessed by CPU [14]. When the memory capacity is reduced, the data access speed needs to be increased, the data is converted into the memory form by the program, and the data is accessed by the CPU through the cache.

2.2. Information retrieval sorting algorithm

The software design adopts an algorithm directly facing the retrieval object, which divides the processing of dynamic information into manual annotation of information, characteristics of information content, basic attributes of information and data sources of information. Manual annotation of information describes the content of dynamic information of the entire remote distributed database in the form of text [15]. The information content is characterized by processing the original information and carrying out corresponding matching operation; the basic attribute of information is to store the type of file, the memory occupied, the date of processing and some other attribute information in the variables of the retrieval object, and to carry out initial editing through the specified program. The data source of information refers to the most original information, and indexes corresponding fields through the establishment of a remote distributed database. The flow of dynamic information loading of remote distributed database is shown in fig. 3.

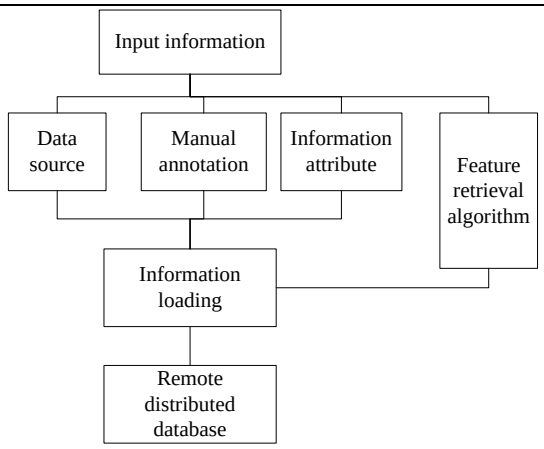

Figure 3. Flow of dynamic information loading

Information loading of dynamic information is an important link. The retrieval system mainly quantifies the human visual perception non-uniformly with different components of the color model, and effectively compresses the dimensionality vectors. Then, the principal component analysis method is used to reduce the dimensionality of the whole information, and the retrieval method to solve the multivariate variance structure through the principal component is studied. The result is the change rate adjacent to it. First, determine the logarithm of the information value:

$$i(a,b)=\ln\left(r(a,b)\right) \tag{1}$$

Where (a, b) is the logarithm of a given point; $r(a,b)$ is the l value of the poin. Then calculate different convolutions:

(1) Laplace convolution:

$$i(a,b)\Delta^2 \Rightarrow c(a,b)$$

$$\Delta^2=\left\{\left[p(a+1,b)-p(a,b)\right]-\left[p(a,b)-p(a-1,b)\right]\right\} \tag{2}$$

$$+\left\{\left[p(a,b+1)-p(a,b)\right]-\left[p(a,b)-p(a,b-1)\right]\right\}$$

(2) Directional derivative

$$i(a,b)\Delta^{n}=c_{n}(a,b) \tag{3}$$

Finally, a histogram is constructed:

(1) Laplace convolution:

$$H_i=\sum_{a,b}W=\begin{cases}1 & c(a,b)=i\\ 0 & c(a,b)\neq i\end{cases} \tag{4}$$

(2) Directional derivative

$$H_i=\sum_{n=1}^{4}\sum_{a,b}W=\begin{cases}1 & c(a,b)=i\\ 0 & c(a,b)\neq i\end{cases} \tag{5}$$

The accuracy rate of the information features obtained from the above algorithm is still relatively high, but at the same time it also produces a higher dimension. Therefore, PCA method must be adopted to reduce the dimension of the information. First set $\{a_i\}^c\ i=1$ is a set of information, where C is the number of several pieces of information in the information set. a_i is the vector of information features. The training set of several samples are randomly extracted from the whole set $\{a_{i1},a_{i2},\cdots,a_{im}\}$, then the expression of the sample's average value n and covariance q is as follows:

$$n=\frac{1}{m}\sum_{j=1}^{t}a_{ij} \tag{6}$$

$$A=\frac{i}{\sqrt{m}}\left[a_{i1}-n,a_{i2}\text{-n},\cdots,a_{im}-n\right] \tag{7}$$

$$n=\frac{1}{m}\sum_{j=1}^{t}\left(a_{ij}-n\right)\left(a_{ij}-n\right)^{T}=AA^{T} \tag{8}$$

Set up $y_i, i=1,2,\cdots,m$ as the eigenvector of covariance Q, The corresponding eigenvalue is β_i, so you can get:

$$Q=AA^{T}y_i=\beta_i y_i \tag{9}$$

According to the above calculation steps, the eigenvalues of covariance q and eigenvectors are obtained, then the calculated eigenvectors are arranged in descending order of magnitude according to the corresponding eigenvalues, and the obtained eigenvalues are transformed into matrices in the form of group columns

$$y_i=Q^{T}a_i \tag{10}$$

Yi in the formula is the feature ranking vector of dynamic information retrieval. Based on the above algorithm, the sorting and pushing of webpage retrieval information can be effectively realized.

2.3. The realization of information retrieval of web page subject

In order to ensure the accuracy of the information retrieval of the main body of the webpage, an interest model is added to the common search program to check the searched results, and then the data of interest to the user is presented to the user through the human-computer interaction interface according to the optimal order. The overall structure of personalized retrieval is shown in the Figure 4.

There are three more functions in personalized search than general search, namely user query agent, searcher and user interest model. The retrieval request input by the user is preprocessed by word segmentation, personalized adjustment, feedback and the like, and the retrieval result is presented to the user according to the preprocessed retrieval request; Matching the user request with the information in the database; By analyzing the retrieval request input by the user, the key point of retrieval, that is, the real interest of the user, is found out, so that the system can provide services in combination with the interest of the user, instead of simply providing services based on retrieval words. Among them, the user interest model is the key to realize personalized retrieval, and it is also the

most exposed part of user privacy information. Therefore, the recommended information is retrieved on the basis of user interest. Therefore, this study focuses on the protection mechanism of the user interest model, that is, to realize anonymity and ensure data privacy security. The process of initial data privacy protection using differential privacy technology is shown in the following Figure 5.

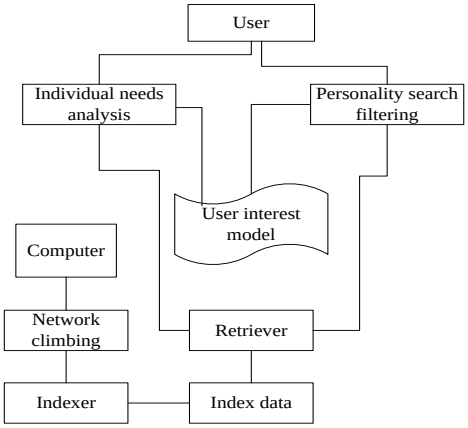

Figure 4. Overall Structure of Personalized Retrieval

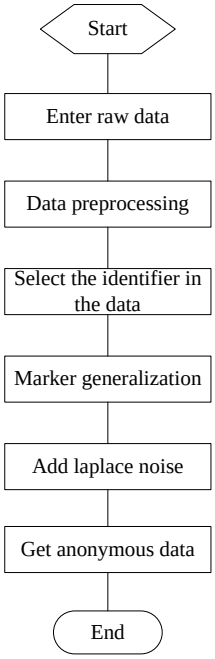

Figure 5. Initial data privacy protection process of differential privacy technology

Based on the differential privacy technology, the primary data privacy protection process extends the generalization of the index indicator, and all the information that is connected with sensitive information is completely captured. The specific retrieval steps are as follows.

Step 1: Sort each attribute value of N identifiers according to the order from top to bottom, and conduct an experimental division between two adjacent attribute values.

Step 2, constructing an interest model equivalence group, and calculating the leakage risk degree of each identifier attribute in the equivalence group under the same sensitive value.

$$A = \frac{a}{n} \tag{11}$$

In the formula, A is the risk leakage ratio; A is the sensitive value; N is the number of tuples in the equivalent group.

Step 3: Calculate the user weight corresponding to the interest model of the equivalent group.

$$U_p(u) = \begin{cases} w, & \text{if}: u \in U_p \\ \dfrac{\sum U_p(u)}{n} \end{cases} \tag{12}$$

Step 4: Compare the weight value with the risk value. If the weight value is greater than the risk value, perform one-level generalization, and then return to Step 2 until the weight value is less than the risk value and generalization is finished.

The higher the generalization level, the stronger the generalization level, but the higher the risk of the data. The most commonly added noise in differential technology is Laplace noise. adding noise will interfere with the attacker's query operation, making the attacker unable to distinguish whether the obtained data is true or false, thus achieving the purpose of protecting the user's privacy data security. The addition of differential privacy technology improves the accuracy of data query, at the same time, it also reduces the probability of identifying individuals and attributes, and reduces the error of data released by the system. Based on the above steps, accurate retrieval of webpage information can be effectively realized.

3. Analysis of Experimental Results

In order to verify the validity of the template web topic information automatic retrieval model based on vector space, comparative experiments were carried out. In order to ensure the experimental effect, the experimental environment and parameters are set uniformly. Specific experimental parameters are as follows:

Table 1. Data privacy protection effect

P value	Retrieval accuracy (%)

	This method (%)	Internet data publishing (%)	k-degree-l-diversity Anonymous search (%)	Convergence function (%)	Time (s)
0.5	80	80	80	80	10
0.6	82	83	82	84	20
0.7	87	85	86	85	30
0.8	90	88	89	86	40
0.9	95	92	91	87	50
1.0	98	95	90	89	60

Carry out comparative tests under the above experimental environment and record the test results for observation and analysis. The specific test results are shown in the following figure:

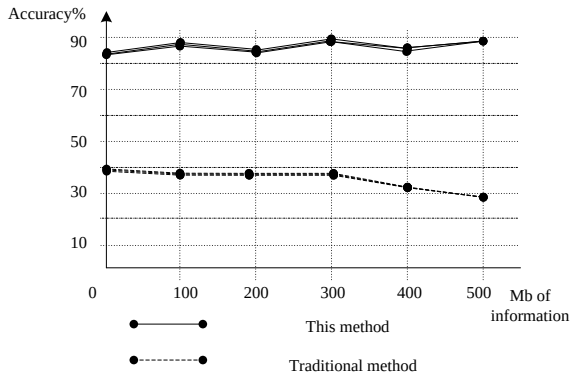

Figure 6. Comparison test results

Observing the above detection results, it can be seen that compared with the traditional retrieval model, the template web topic information automatic retrieval model based on vector space proposed in this paper has relatively higher retrieval accuracy effect in the actual application process, has higher practicability under mass data, and fully meets the research requirements.

4. Conclusion

This paper probes into three methods for automatically evaluating the information retrieval system in the question and answer system. They are classification, regression and sorting. In order to improve the accuracy and timeliness of data retrieval in the massive information environment, 12 indexes describing the correlation between the question itself and the answer to the question and the document are mainly studied, and the template web topic information automatic retrieval model is optimized by combining the vector space principle.

References

[1] Liu Menglan, Liu Bin, Peng Zhiyong. Automatic patent query expansion based on word embedding. Computer Engineering and Science. 2017, 39(12), 2297-2305.

[2] Shen Qiangqiang, Xiong Zeyu, Xiong Yueshan. A new automatic summarization method based on paragraph vector. Computer Engineering and Science. 2019, 24(6), 1064-1070.

[3] Chen Pei, Xu Lei, Liu Kun. Research of electrical appliences product customization case retrieval method. Modular Machine Tool & Automatic Manufacturing Technique. 2017, 26(2), 101-103.

[4] Wu Yongliang, Chen Jianping, Jia Zhijie, et al. The construction of geological data ontology and its application to data retrieval. Geological Bulletin of China. 2018, 37(5):945-953.

[5] He Xing, He Huan, Dong Chongchong, et al. Design of automatic monitoring system for intelligent information security protection in cloud environment. Electronic Design Engineering. 2019, 27(10), 169-172.

[6] Wu Yirui, Huang Duohui, Zhou Yiyang. Design and implementation of water information retrieval system based on web crawler. Water Resources Informatization. 2017, 21(4), 36-41.

[7] Li Yuelin, He Pengfei. A user study on gamified information retrieval systems: preference for game elements, attitude, and usage intention. Journal of Library Science in China. 2019, 45(3), 62-78.

[8] Li Hui. Research on the Intelligent Retrieval System of Library Based on Semi-structured Collection data. Journal of Library and Information Sciences in Agriculture. 2017, 29(2), 61-64.

[9] Sun Min, Ju Yong. Design of hydraulic literature retrieval system based on SSM and Lucene. Water Resources Informatization. 2019, 25(1), 19-24.

[10] Huang Xiaoqin, Dai Jingjuan, Xu Haidong. Research of intelligent full-text retrieval system of electronic medical record based on big data hbase. China Digital Medicine. 2019, 25(5), 27-29.

[11] Wang Peiguang, Wang Peng, Fan Yiping. Research on disk retrieval system of mobile terminal operating system. Technology Wind. 2017, 24(2), 77-78.

[12] Guo Chaonian. Lucene test case searching system based on banking keywords. China Computer & Communication. 2017, 14(8), 140-142.

[13] Chen Lin. Design of intelligent information retrieval system under big data environment. China Computer & Communication. 2017, 24(6), 128-129.

[14] Gu Shen, Gu Shen. Research on library information retrieval service system based on distributed structure. Modern Electronics Technique. 2017, 40(1), 83-85.

[15] Liu Runzhen, Chen Tao, Chen Honghui, et al. Application of OLAP analysis and intelligent supporting retrieval based on bidding data in bidding system. Computer and Digital Engineering. 2018, 46(7), 1481-1486.