International Journal of Intelligent Information and Management Science
ISSN: 2307-0692, Volume 8, Issue 3, June, 2019

Optimization of Heterogeneous Information Real-Time Retrieval Model of Campus Website

Zhiyong Chen
Education Technology and Information Center, Guangdong Medical University, Zhanjiang, 524023, China

Abstract: Most of the current information retrieval models of campus websites are centralized. In the case of large amount of information or more requests for access at the same time, it reduces the real-time performance of the retrieval and causes the bottleneck of the network. To address this problem, a new heterogeneous information real-time retrieval model of campus website is designed in this paper. The model of heterogeneous information network is introduced. LDAP directory technology is used to organize and manage global index information. A distributed heterogeneous information retrieval model of campus web site based on LDAP is designed and its physical architecture is built. The retrieval node is designed by using distributed structure of network token-ring and LDAP protocol. The queue is retrieved by using Pop Rank method and Co-Hits method. A large number of related heterogeneous information will be returned in the case of large difference between the retrieval key words and the required retrieval results, which need to be optimized. Rough classification of the retrieved results is achieved by clustering method to remove the result that is far apart from most of the retrieval results. The SOM neural network is used for fine classification to close the desired results and improve the real-time performance of retrieval. Experimental results show that the designed model is highly accurate and real-time.

Keywords: Campus website; Heterogeneous information; Real-time; Retrieval model; Optimization

1. Introduction

The number of information in the campus website is large, and the variety and structure of the information are more and more. This brings difficulties for students and teachers to obtain the required information. The information resources in the information space of campus website are strong. Structured information, semi-structured information, and unstructured information exist together in the campus website. The storage systems and operating systems of the information sources are different. The information source is stored on the various sites and has different degree of autonomy. In addition, each storage site will also add, delete, or modify information according to the actual needs, causing the information source with a certain dynamic change.

At present, most of the current information retrieval models of campus websites are centralized. In centralized information retrieval system, information resource is first retrieved, the index information is stored on the centralized server, and all users' queries are sent to the same server Choi. Although this method is simple, it will not only reduce the efficiency of the system, but also cause the bottleneck of the network when there is a large amount of information or more requests at the same time. Distributed information retrieval can create an effective index structure and provide good scalability and retrieval performance, which is widely used in heterogeneous information retrieval. In the existing information retrieval system, metadata is used to describe the website resources of the campus website, and then, on the basis of metadata information, it provides users with information retrieval services. In this paper, based on the background of the heterogeneous information of campus website, the information retrieval model based on LDAP directory service is proposed and optimized. LDAP protocol is the main protocol to support directory services. It has a good hierarchy and scalability, supports multiple data types, and can achieve access control. In addition, it has distributed characteristics, which can manage physically distributed metadata uniformly, while maintaining the logical consistency and integrity of these metadata. Therefore, it is suitable for organizing and managing metadata in heterogeneous information, so as to provide information retrieval services for users.

2. Optimization of Heterogenous Information Real-Time Retrieval Model of Campus Website

2.1. Heterogeneous information network model

A general information network is an information network that contains one type of information or does not distin-

guish the type of data, also known as isomorphic information network. When the type of the information object is two or more than two, the information network is called the heterogeneous information network.

Define $G=(U,L,Q)$ is a heterogeneous information network. When the node set satisfies $U=R_1\cup R_1\cup R_n$, The nodes in U belong to n different class of information, where $R_i(1\le i\le n)$ is the object set of the class i. L and Q are set of the edge G of the heterogeneous information network and the corresponding weight matrix, respectively. The topology of heterogeneous information network is shown in Fig. 1.

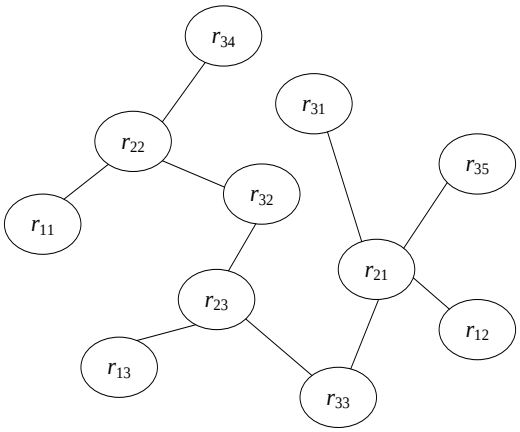

Figure 1. Topology of heterogeneous information network

In Fig. 1, r_{ij} represents the j th object of R_i type object. The network nodes are composed of three different objects, which are $R_1(r_{11},r_{12},r_{13})$, $R_2(r_{21},r_{22},r_{23})$, and $R_3(r_{31},r_{32},r_{33},r_{34},r_{35})$. The three types of nodes are connected in accordance with a certain relationship. From the relationship diagram, objects of two types of R_1 and R_3 are not directly associated, but connected indirectly through the objects of the R_2 type. The matrixes Q_{21} and Q_{23} represent the weights of the association relationship of R_2-R_1 and R_2-R_3, then $Q=(Q_{21},Q_{23})$. Q_{21} and Q_{23} are expressed as

$$Q_{21}=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix} \quad (1)$$

$$Q_{23}=\begin{bmatrix}1&0&1&0&1\\1&0&0&0&1\\0&0&1&0&1\end{bmatrix} \quad (2)$$

The relationship between the two objects is symmetric without considering the directionality. Therefore, $Q_{21}=Q_{12}^{T}$ and $Q_{23}=Q_{32}^{T}$, that is, the relation matrix is a symmetric matrix.

For heterogeneous information network, the relation weights between different types of network nodes are generally different, which is mainly determined by the role of different types of nodes in different applications.

2.2. Design of heterogeneous information retrieval model based on LDAP

The lightweight directory access protocol (LDAP) standard was developed by University of Michigan in July 1993, which is the simplification and improvement of the X.500 standard, and provides TCP/IP protocol support for the DAP protocol in X.500. There are two versions of the currently used LDAP protocol, which are LDAPV2 and LDAPV3. LDAP is a cross platform protocol, and applications with LDAP are applicable to any server. The LDAP directory service uses the tree structure. LDAP has four basic models and its servers are easy to build and maintain and have good security.

LDAP is a hierarchical database of data models. The information is organized according to the tree structure, and the directory is made up of entry. Entry corresponds to record in the relation database. Entry is an attribute set with distinguished name (DN). The top layer of the LDAP directory tree, the root of the directory tree, is the base of directory tree, also known as base distinguished name (Base DN). Most LDAP directories use organization unit (OU) to logically separate the data. The record item stored in the LDAP directory has a name, which is usually saved in the common name (CN) attribute. DN is equivalent to primary key in the relation database for reading the record. Attribute is made up of type and multiple values, equivalent to a filed in a relation database, which consists of filed name and data type. In LDAP, the organization of entry is usually organized according to geographical location and organization relationship. Data is stored in files. In order to improve efficiency, index-based file database can be used instead of relation database. The DAP supports the attributes that the entry must contain, which is implemented by a special attribute called the object class. The value of this attribute determines some of the rules that the entry must follow.

From the above analysis, the LDAP directory service has the features of openness, distribution, scalability, and cross platform. Depending on the query speed and good performance of the LDAP directory server, the ability of heterogeneous information retrieval on the campus website will be greatly enhanced.

2.3. General framework and function description of heterogeneous information retrieval model

The heterogeneous information retrieval on campus website is completed on the basis of information organization and index. Based on the analysis of the current development of information retrieval and the characteristics of heterogeneous information, LDAP directory technology is used to organize and manage global index information, and a distributed heterogeneous information retrieval model based on LDAP is designed.

The model includes three layers, which are information resource layer, information service layer, and user interface layer.

Information resource layer. It is the bottom of the retrieval model. It is composed of various information resources distributed throughout the information space, including text information, hypertext information, and audio and video information. These information resources are organized and managed by the campus network database management system as a storage medium. Information resources are registered and entered into the metainformation storage of the local information resource center after the packing of the packer, while some commonly used information in information sources can be stored in the resource center for local information storage after proper conversion.

Information service layer. It is the core of the retrieval model, including LDAP directory service, resource center information management and service, query decomposition and query result integration. The LDAP directory service is used to organize and manage global index information. The information in the directory database consists of two parts: local index information and global index information. LDAP directory servers are organized in a hierarchical organization.

After receiving a query request from a user, the directory service is queried first. When the location of the resource center for the required information is obtained, the request is forwarded to the resource center for execution. If the required information can be found in the local repository of the resource center, the query results are returned to the user. Otherwise, according to the location of the specific information source provided by the resource center, the query request is decomposed to the information source for execution.

Information presentation layer. This is the highest level of the retrieval model and is responsible for the interaction with the user. The layer can receive information request from the user and return the request response to the user in an appropriate form.

2.4. Retrieval node design

Distributed structure design of retrieval node is designed by using LDAP protocol and network token-ring. The retrieval nodes can be deployed on multiple servers. Each node communicates through a virtual heartbeat mode and reports the health and load in real-time. When the failure or overload of a retrieval node is retrieved by host process of retrieval queue engine, it will actively assign the next search task to the lighter load retrieval node.

For each worker process, the independent process design makes the worker not need to lock when handling the task and save the cost of the lock. The use of the independent process allows each retrieval task to not affect each other. After an unusually exiting of a process, the master process will quickly start the new worker process to replace the work of the exception exit process. The setup of this exception mechanism will cause all requests on the worker to fail, but it will not affect other requests, so it reduces the crash risk of the whole search engine.

2.5. Design of queue characteristics of retrieval model

The design of the retrieval queue is a key technique for the heterogeneous information retrieval model of campus website. In this paper, preemption technology is used. The higher priority retrieval task of the user is first guaranteed to execute. The implementation is shown in Fig. 2.

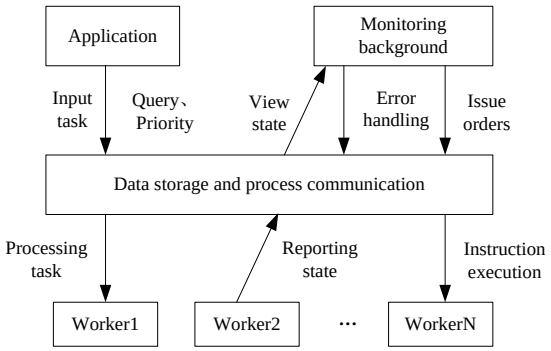

Figure 2. Principle of retrieval work

In order to sort nodes on campus websites, a series of sorting methods are proposed. The representative methods include Pop Rank method and Co-Hits method.

In Pop Rank method, the web page of the campus network is regarded as a combination of various types of objects, and the page is transformed into a heterogeneous information network. For various types of links in the campus website, Pop Rank assigns different popularity propagation factors (PPF) to different types of connections. A random object finder model is proposed. The model believes that the user's goal is to find the object information in the campus website. So it chooses the page that contains the object according to the popularity of the page, and chooses according to the relationship between the heterogeneous object and this object. The sorting model is given by:

$$D_X = \delta D_{HX} + (1-\delta)\sum_{\forall Y} v_{YX} K^T_{YX} D_Y \quad (3)$$

where $X=(x_1,x_2,\cdots,x_m)$ and $Y=(y_1,y_2,\cdots,y_m)$ are two types objects, D_X and D_Y is the popularity vectors of X and Y, D_{HX} is the popularity vector contributed by the page of the object X, δ is the damping coefficient, v_{YX} is the transfer factor of information between heterogeneous objects and $\sum_{\forall Y} v_{YX}=1$, $K^T{}_{YX}$ is the adjacency matrix representing the relationship of X and Y, which is given by

$$\begin{cases} k_{yx}=\dfrac{1}{N_{um}(y,X)}, \text{There is a link between y and X} \\ 0, \text{else} \end{cases} \quad (4)$$

where $N_{um}(y,X)$ is the sum of the number of the links from y to the nodes of X.

Co-Hits models bipartite graph, which is the most common model in campus web sites with heterogeneous information. The relationship of heterogeneous nodes is transformed into the transition probability between the homogeneous nodes for analysis. For $G=(U,L,Q)$, $u_i \in U$, $l_k \in L$, x_i is the ranking score of u_i, y_k is the ranking score of l_k, which satisfy

$$x_i=(1-\sigma_u)x_i^0+\sigma_u\sum_{k\in L}p_{ki}^{lu}y_k \quad (5)$$

$$y_k=(1-\sigma_l)y_k^0+\sigma_l\sum_{i\in U}p_{ik}^{ul}x_i \quad (6)$$

where $\sigma_u \in [0,1]$, $\sigma_l \in [0,1]$, x_i^0 is the initial ranking score of u_i, y_k^0 is the initial ranking score of l_k and $\sum_{i\in U}x_i^0=1$, $\sum_{k\in L}y_k^0=1$, p_{ki}^{lu} is the transition probability from l_k to u_i, p_{ik}^{ul} is the transition probability from u_i to l_k, which can be obtained by adjacency matrix in bipartite graph. The method uses the connection between the heterogeneous nodes to enhance each other, and can obtain more reasonable results than the initial ranking.

3. Real-Time Optimization of Retrieval Model

For the designed heterogeneous information retrieval model of campus website, the database is the basis of retrieval, which provides the entry point of knowledge acquisition. In campus websites, some users are more aware of the retrieved heterogeneous information, so the keywords they provide are generally more professional. But for users who are not particularly aware of heterogeneous information, the keywords they enter are different. The retrieval model retrieves all the heterogeneous information related to the keyword input by the user. The retrieved data must be further filtered to eliminate the erroneous and inaccurate results, and obtain the most satisfactory retrieval results, so as to achieve the purpose of intelligent retrieval.

In this paper, two methods of clustering and neural network are used for classification and comparison. First, the clustering method is used for rough classification of the retrieved results, which remove the most unlikely results and reduce the scope of the retrieval results. Then according to the obtained results, the SOM neural network method is further used for fine classification to obtain the desired results.

3.1. K-means clustering method

The K-means clustering method was first proposed by Mac Queen in 1967. It is a basic division method in clustering method, and sum of squared error criterion is often used as a clustering criterion function.

The principle of K-means method is as follows. First, K points are selected randomly from the data set as the initial clustering center. The distance from each sample to the cluster is calculated. The sample is classified into the class in which the nearest cluster center is located. The new clustering center is obtained by calculating the average value of each newly formed data object. If there is no change in the two adjacent clustering centers, it shows that the sample adjustment is over, and the clustering criterion function has been converged.

K-means clustering method is used for rough classification of the obtained heterogeneous information after retrieval. The idea is as follows. Give a set with n_y elements, The set is divided into K clusters and $K<n_y$. K clusters satisfy the following condition. Each cluster contains at least one element and each element belongs to and only to a cluster.

For the given K, first, an initial grouping method is presented, which makes the optimization degree of the grouping scheme decrease compared with the previous optimization after each improvement, while the element diversity increases in different groups. Then revise through iteration method.

3.2. Fine classification optimization with self-organization feature mapping neural network

After rough classification with K-means, a fine classification must be carried out to achieve more accurate classification and obtain the correct result. Artificial neural network is suitable for solving this problem. In this paper, the used neural network is self-organization mapping (SOM). The structure of SOM model is shown in Fig.3.

In Fig.3, the neural network consists of input layer and competitive layer. The input layer is composed of Ns input neurons. The competitive layer is composed of Ms input neurons. They form a two-dimensional planar array. The neurons of the input layer are connected with the neurons of the competitive layer. There is no lateral inhibition connection between the neurons of the competitive layer.

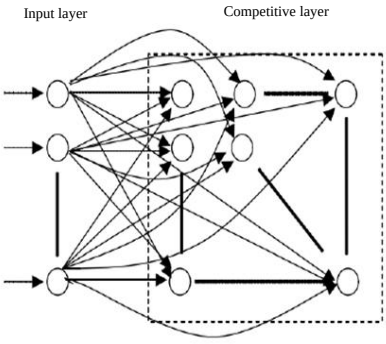

Figure 3. SOM neural network structure

Through repeated learning of input pattern, the pattern features of each input mode are found and organized, and the classification results are showed in the competitive layer. Compared with other neural network, the features are as follows. It does not express the result of classification based on a single neuron or a single neuron vector, but it results from simultaneous classification of several neurons in the neural network.

4. Experimental Results and Analysis

4.1. Experimental data set

In order to evaluate the designed retrieval model, data collection in campus website not only includes data of text and category, but also needs click data on these texts and categories. However, there is no such standard test set at present, so it must be built manually. The used original dataset is the click record from some campus website in December 2017. The original format of the dataset is as follows.

Each click record contains the information of time, user ID, query term, URL's ranking in the return result, and the clicked URL. The data set consists of 51044526 log records, 5737315 different queries, and 15962315 different documents. In order to clear the noise data in the record, obtain different category of click information, URL-based classification directory provided by the campus network website is used for processing. The obtained data of five categories is shown in Table 1.

Table 1. Experimental data category statistical information

Category	Category 1	Category 2	Category 3	Category 4	Category 5
Document	12335	29865	13826	34492	16285
Query	3936	7822	4351	9178	4596
Average clicks	12.28	12.88	10.07	15.79	13.52

4.2. Performance evaluation

In this paper, accuracy (AC) and normalized mutual information (NMI) are applied to evaluate the retrieval performance of different models. AC is defined as

$$AC = \frac{\sum_{i=1}^{e} b_i}{n_z} \tag{7}$$

where e is the number of categories, b_i is the number of correctly retrieved samples, and n_z is the total number of samples.

NMI is defined as:

$$NMI = \frac{\sum_{s=1}^{e}\sum_{w=1}^{e} n_{s,w} \log\left(\frac{n_z n_{s,w}}{n_s \hat{n}_w}\right)}{\sqrt{\sum_{s=1}^{e} n_s \log\left(\frac{n_s}{n_z}\right)}\sqrt{\sum_{w=1}^{e} \hat{n}_w \log\left(\frac{\hat{n}_w}{n_z}\right)}} \tag{8}$$

where n_s is the number of samples in the s th category, $\hat{n}_w$ is the number of samples in the w th category in the retrieval results, and $n_{s,w}$ is the number of samples in the standard result category e_s and the retrieval result category $\hat{e}_w$ at the same time. The more consistent the results are with the standard results, the greater the NMI value is, and the NMI is l when the two results are completely overlapped.

In the experiment, first, only one type of data is labeled. 1%, 2%, 3%, 4%, 5% of the retrieved data from each type of retrieval result is randomly extracted. Then the heterogeneous information is retrieved on the campus website by using the proposed model, ontology retrieval model, and LDA retrieval model. 5 experiments are carried out on each ratio of sample, and the means of the results are obtained. The results are shown in Table 2.

Table 2. Comparison of retrieval performance

Sample percentage /%	The prosed model		Ontology model		LDA model	
	AC/%	NMI/%	AC/%	NMI/%	AC/%	NMI/%
1	78.26	69.35	63.35	42.86	53.85	32.65
2	76.35	72.66	66.51	43.71	54.5	33.42
3	79.61	75.13	65.82	42.56	54.69	33.8
4	82.16	76.75	68.31	45.22	59.2	38.56

5	75.03	78.2	69.52	47.96	63.6	45.1

From Table 2, it can be seen that, the accuracy of the proposed model is much higher than ontology retrieval model and LDA retrieval model. This is mainly because the classification features provided by heterogeneous information are not significant. It cannot be effectively retrieved with ontology retrieval model and LDA retrieval model.

4.3. Real-time test

In order to verify the advantage of the proposed model in real-time, the response time of the proposed model, ontology model, and LDA model is tested with different keyword and sample number as input. The statistical results are shown in Fig. 4. From Fig. 4, it can be seen that, in the case of increasing number of key words, the response time of the proposed model has no significant change, and is always lower than the ontology model and the LDA model, which shows the better real-time quality.

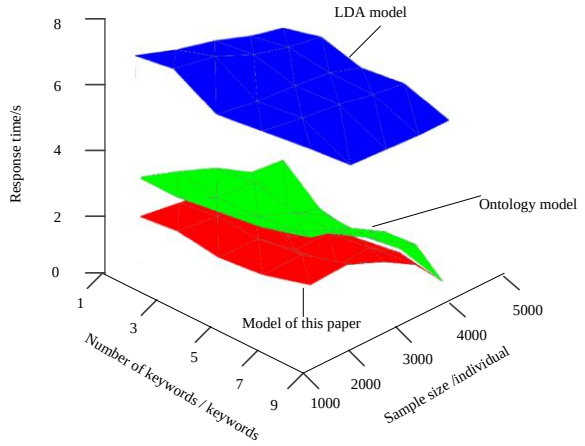

Figure 4. Real-time test results

5. Conclusions

In this paper, a new real-time retrieval model of heterogeneous information on campus website is designed. The model of heterogeneous information network is introduced. A distributed heterogeneous information retrieval model of campus website based on LDAP is designed. The design process of retrieval node and retrieval queue is given. For the drawback of the design model, rough classification with clustering and fine classification with SOM neural network are used for optimization to obtain the desired retrieval results and real-time retrieval. Experimental results show that the designed model is highly accurate and real-time.

References

[1] Arour K., Yeferny T. Learning model for efficient query routing in P2P information retrieval systems. Peer-to-Peer Networking and Applications. 2015, 8(5), 741-757.

[2] BöHm T., Klas C.P., Hemmje M. Collaborative Information Seeking and Retrieval in a Heterogeneous Environment. Computer. 2014, 47(3), 32-37.

[3] Brosseau-Villeneuve B., Nie J.Y., Kando N. Latent word context model for information retrieval. Information Retrieval. 2014, 17(1), 21-51.

[4] ChoiChoi S., Choi J., Yoo S., et al. Semantic concept-enriched dependence model for medical information retrieval. Journal of Biomedical Informatics. 2014, 47(2), 18-27.

[5] Cummins R., Paik J. H., Lv Y. A Pólya Urn Document Language Model for Improved Information Retrieval. Acm Transactions on Information Systems. 2015, 33(4), 1-34.

[6] Li J.Q., Yang J.J., Liu C., et al. Exploiting semantic linkages among multiple sources for semantic information retrieval. Enterprise Information Systems. 2014, 8(4), 464-489.

[7] Li Xinwei. Optimization Simulation of Information Interactive Efficiency in Layered Heterogeneous Network . Computer Simulation. 2017, 34(1), 276-279.

[8] Moulin C., Largeron C., Ducottet C., et al. Fisher Linear Discriminant Analysis for text-image combination in multimedia information retrieval. Pattern Recognition. 2017, 47(1), 260-269.

[9] Mutschke P., Mayr P. Science models for search: a study on combining scholarly information retrieval and scientometrics. Scientometrics. 2015, 102(3), 2323-2345.

[10] Pecina P., Dušek O., Goeuriot L., et al. Adaptation of machine translation for multilingual information retrieval in the medical domain. Artificial intelligence in medicine. 2014, 61(3), 165-185.

[11] Singh J., Sharan A. A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Computing & Applications. 2014, 28(9), 1-24.

[12] Sumiya K., Kitayama D., Chandrasiri N.P. Inferred Information Retrieval with User Operations on Digital Maps. IEEE Internet Computing. 2014, 18(4), 70-73.

[13] Thangaraj M., Sujatha G. An architectural design for effective information retrieval in semantic web. Expert Systems with Applications. 2014, 41(18), 8225-8233.

[14] Wu Zhenxing, Zeng Lingwei, Wang Wenbin. Optimization of Multi-source Information Retrieval Based on Concept Lattice Feature Partition. Bulletin of Science and Technology. 2015, 31(8), 174-176.