The Development and Application of Chinese Intelligent Question Answering System Based on J2EE Technology

Shouning Qu
School of Information Science and Engineering
University of Jinan
Jinan, China
qsn@ujn.edu.cn

Bing Zhang
School of Information Science and Engineering
University of Jinan
Jinan, China
tanghulu116@163.com

Xinsheng Yu
Shaoxing Electric Power Bureau
Zhejiang Electric Power Corporation
Zhejiang, China
xinshengyu2007@163.com

Qin Wang
School of Information Science and Engineering
University of Jinan
Jinan, China
wangqinjida@yahoo.com.cn

Abstract

The current status of Chinese question answering system was introduced in the paper firstly, some defects of it were pointed out, then a set of scheme about data warehouse design based on data mining was put forward. It applied the improved association rules algorithm to text clustering algorithm. At the same time in combination with the advantages of J2EE architecture in system development, this paper described how to apply J2EE platform to forming an exact and efficient intelligent question answering system by the technologies such as JSP, Servlets, EJB, JDBC and so on.

1. Introduction

Along with the development of computer, network and modern education methods, various teaching information are accumulated, the proportion between students and teachers is becoming lager and lager, all of these make the mode of teacher answering the student's questions face to face become less and less practical. Many question answering systems on educational website now usually use message board, e-mail, BBS, chatting system or Blog to process the question answering, but the message board, e-mail, BBS or Blog can't answer the question in time, while the mode of chatting system requires some teachers to answer the students' large numbers of questions on-line at any moment. Upon that, an intelligent automatic question answering system is demanded urgently. As an important part of implements used to communicate between the teachers and the students, Intelligent Question Answering System (IQAS) is not only the useful supplement to systematic study, but also an important way to consolidate knowledge for students. At the same time, the system can point out the weak parts of student's knowledge by analyzing the track records of the questions asked by the students, so it becomes an effective implement to help teacher to ameliorate the way of teaching.

At present, most of question answering systems are based on keywords finding. The flow of this method is described as follows: users input the keywords firstly, then matching with the Q-A pairs in background database by the system querying function. But as this method uses the technology based on keywords matching, it doesn't relate to the problem that how to analyze the questions by naturally language comprehension. So it demands the user should have the ability of drawing out the keywords accurately. If the keywords input are not precise enough, it may return many questions which are irrelevant to users. Therefore, this kind of question answering system is fast at speed, but low at veracity. Upon that, a set of scheme about the question answering system based on data mining is put forward. An improved association rules algorithm is applied to text clustering algorithm, by this way the system could draw out questions intelligently and update the answers automatically. At the same time, J2EE technology is used to develop the system to make up the defects of current Chinese question answering system. Thereby, the students and teachers could use the system more expediently, the administrator could manage the system more easily, and then an efficient accu-

e-Forensics 2008, January 21-23, 2008, Adelaide, Australia.

rate intelligent question answering system based on limitative field is realized finally.

2. Relevant algorithm and key technology

The performance of intelligent question answering system is mainly evaluated by its accuracy and speed. The tolerable time of users waiting in front of the browser is generally limited, so the speed of the system responding to the users should be fast. The algorithms such as word segmenting, quickly orienting and question matching are applied in this system. On the other hand, users are usually anxious to receive the accurate answers, so the accuracy of system should be high.

2.1. The improvement and application of association rule algorithm

Along with the increase and cumulation of vast data, association rule has become an important aspect to the field of data mining. Apriori association rule mining algorithm is one of the classical algorithms, but it has some defects. According to these defects, the paper puts forward an improved association rule algorithm. The details are described as follows:

Input: Database D of transaction; minimum support threshold min-sup.

Output: the frequent sets L of D.

$L_1 = find\_frequent\_1 - itemsets(D)$

k=1

do while($L_{k-1} \neq \emptyset$)

k++

$find\_frequent\_itemsets(min\_sup)$

move record from D where itemnum=k-1

loop

end

In order to compress C_k, the Apriori algorithm uses the prune step. The prune codes are described as follows:

Procedure $has\_infrequent\_subset(c : candidatekitemset; L_{k-1} : frequent(k-1)\_itemset)$;

2.2. Text clustering

As the text clustering here is the classical short document clustering, the paper puts forward an improved k-means clustering algorithm. The detailed process of this algorithm is shown in Figure 1.

Firstly, the documents should be processed by the Chinese word segmentation technology, which is used in the paper is Maximum Matching Method.Chinese word segmentation technology belongs to the category of natural language processing technology [5]. The flow of the thought is

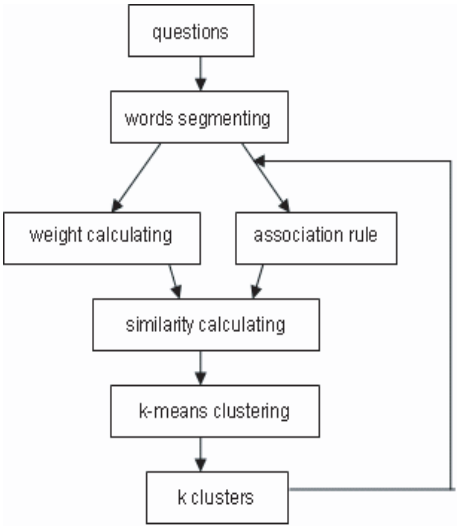

Figure 1. Flow chart of clustering

described as follows: calculate the length of the document firstly, choose n characters from the left to right within the range of the document's length, here n should be determined by the length of the longest word in the background words table. Then match the n characters with the words in table, judging it matches successfully whether or not. If it dose, segment a word and cut the word from the documents; if not, subtract a character from the right of the n characters, continue matching. If it doesn't match successfully until there is only one character left in the n characters, it implies that this word isn't contained in the table. Next, begin from the second character of the document, choose n characters again, and then continue matching with the words in table. If it matches successfully, the next step should begin from the character after the last character of the word. The word segmentation algorithm is applied to every document and the results are saved in a table. The following step is separated into two parts: on the one hand, weight calculation is applied to every word of the document, considering the effect of the length of the document to weight, the paper uses TFC calculation method which is normalizable. The formula is shown as (1).

$$weight(i) = \frac{TF_{kj} * \log(D/DF_i)}{\sqrt{\sum_{k=1}^{M}[TF_{kj} * \log(D/DF_k)]^2}} \quad (1)$$

In the formula, TF_{ij} is the frequency of the key word i in document j. D is the number of all documents. DF_i is the number of documents that have the key word i. Then use the Vector Space Model to denote the text eigenvector by weight.

On the other hand, all documents are regarded as transaction database, each document (question or answer) is re-

garded as a transaction, the key words in documents are regarded as a group of items. In this way, the problem of association analyzing between the key words in text database turns to the association mining between the items in transaction database. At the same time, the problem of similarity analyzing between the questions turns to be the problem of association rule between the items in transaction database. It calculates the correlation value between key words in documents to form the association matrix. Here, when calculating the correlation between key words, the association front and rear are all written in the association table, the format of the association table is described as(no front , rear, S, C), the meanings of it is (the serial number of association rule, association front, association rear, support, confidence).

2.3. Similarity computation

The paper computes the similarity degree based on association rule. It assumes that one document in the database is composed by m words (W_1,W_2...W_m), another one is composed by n words (W_1,W_2...W_n), then the relationship matrix between the two documents can be expressed as follows:

R_{11} R_{12} R_{13} ... R_{1m}
R_{21} R_{22} R_{23} ... R_{2m}
... ...
R_{n1} R_{n2} R_{n3} ... R_{nm}

R_{nm} is the correlation value between the keywords W_n and W_m. The detailed algorithm is described as following:

if W_m=W'_nthen R_{nm}=1
else for each record of association table
if W_1 and W_2 record then i++;
R_{nm}=(i-a)/(i +a) // a is a default parameter
else R_{nm}=b // b is a default parameter

The value of a (or b) lies on the numbers of databases. They are generated randomly at first and finally determined through testing repeatedly.

After calculating the correlation value between key words, the similarity degree between sentences can be computed. The weight of document 1 (W_1,W_2...W_m) can be described as (T_1, T_2...T_m), and the weight of document 2 (W_1,W_2...W_n) can be described as (T'_1, T'_2...T'_n). The algorithm of similarity degrees calculation between the two documents shows as follows:

S_i=S'_i=sum1=sum2=0;
For ($i = 1; i \leq n; i++$)
For ($j = 1; j \leq m; j++$)
S_i= S_i +T'_i* R[i][j];// S_i is the correlation value of W'_i and the document 1
S'_i= S'_i + T_i* R[i][j];// S_i' is the correlation value of W_i and the document 2
sum1=sum1+S_i; sum2=sum2+ S'_i; //end i
sim=(sum1+sum2)/2;//sim is the similarity degree between the document 1 and the document 2

After computing the similarity degree, the clustering algorithm can be executed to gain some cluster of subsets, then the association rule algorithm is executed on every subset, viz. the clustering algorithm is used as the pretreatment of association rule. In this way the correlation value between words of every subset can be obtained, so t he relationship matrix could be more exact. This step can be executed for n times to improve the veracity of clustering.

3. Structure design of system data warehouse

The process of the data warehouse mainly includes three steps: marshal all kinds of original data, manage the data and obtain the information which is needed [7]. The data of this question answering system is come from the background information database of a campus question answering forum and some information on other question answering forums. The original format of the information is one question corresponding more than one answers, furthermore, here the differences among these answers are not cared. Then array these answers according the similarity degrees between them, the best answer is the one which has the biggest similarity degree. In this way the Q-A pairs which are processed simply are formed. Organizing the database by k-means text clustering algorithm to classify the Q-A pairs, so when users raising a question, it could be ranged to a certain field fleetly by the classified data above, then users could get the best answer and the reference answers from this certain field, in this way the speed of answering questions can be advanced consumedly.

As most of questions raised by students are similar, if the Q-A warehouse is designed successfully enough at the beginning, lots of questions can be included in, the system could return the interrelated answers in time, so the workload of teachers could be reduced greatly. After the system using for period of time, the Q-A warehouse will be extended and updated automatically due to the new or better answers, the practicality of the system will be better and better. The flow chart of the data warehouse formation is as figure 2.

4. Implementation of system J2EE four tiers

The system's main function structure is as figure 3.

The classical J2EE application structure includes four tiers: Client tier, Web tier, Business tier and Enterprise information system tier (EIS tier). The second tier and the third tier here are called by a joint name Middle tier. Client tier can be web browser or desktop applications; Web tier and Business tier are deployed on Application Server, implementing by some normative J2EE groupware such as

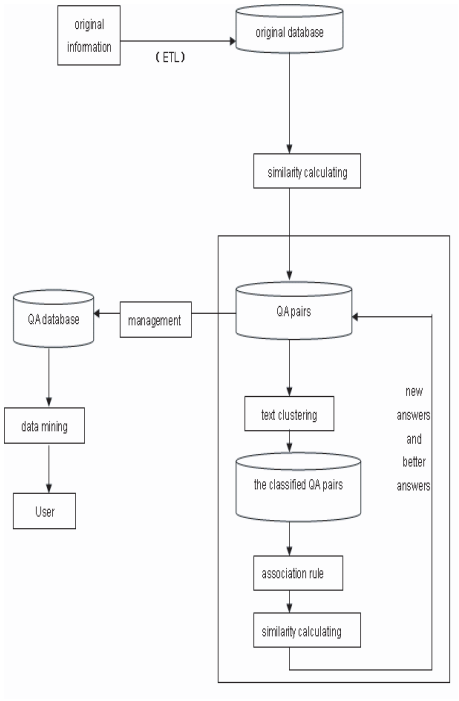

Figure 2. The formation of data warehouse

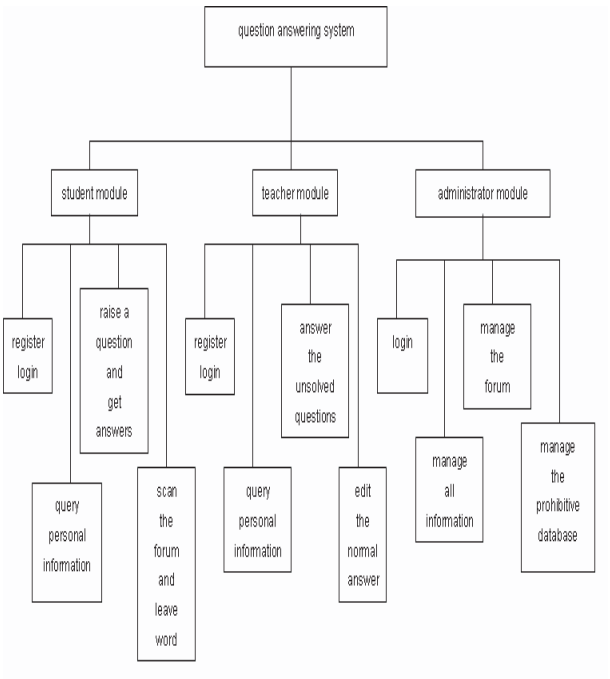

Figure 3. System function structure

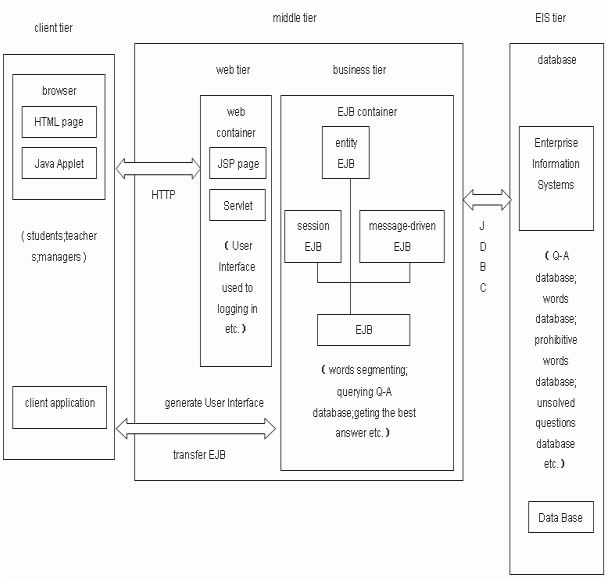

Figure 4. J2EE structure of question answering system

JSP, Servlet, EJB etc. EIS tier is mainly applied to the management of enterprise information, J2EE application groupware accessing EIS tier to obtain some data information frequently [4]. Then the paper narrates the technologies provided by J2EE which are applied to the system in detail as follows. The J2EE structure figure of system is as Figure 4.

Client tier applied to this system is composed of two parts: one is the Client based on web browser, it orients the numerous students who are dispersive and change frequently, the function of this part is difficult to deploy by the conventional Client; Besides, the privilege of students is limited to raising a question and submitting the questions which are not included in Q-A database to the unsolved question database. It doesn't relate to the matter such as modifying the system information directly, so it affects the system security little by network. The flow is described as follows: users start up the browser at Client, connect with web server which could create the dynamic HTML information by networking, then transmit the service required by users to Business tier to do some interrelated analysis, finally obtain the information of the question by querying the database in EIS tier. The system will return the information in the format of web pages to users, i.e. users input the question at Client, the question will be processed by question answering server, the interrelated answers can be returned by querying the system database at last. The other one is the special Client based on C/S pattern, it orients the teachers and administrators. Applying the special Client to system maintenance and management is mainly for the sake

of system security.

The design of Presentation layer based on Web Client mainly includes the register JSP and login JSP. The paper applies the MVC pattern which is popular in web applications to develop this part of the system. Here MVC is Model-View-Controller. In this pattern, View is the interface used to communicate among users [2]. The application of it in this system is some dynamic JSP pages, such as the interface used to login, the interface used to raise a question and so on. These interfaces are deployed in Web Server. Model is the main body of the application, it is used to express business data and business logic. A Model can provide data for multi-view, so the reusability of codes can be improved. The application of it in this system is the EJB groupware (the primary groupware is Entity Bean) running in EJB container, including Q-A entity, student entity, teacher entity, administrator entity etc. In addition, all kinds of data manipulation limited in these entities could be regarded as a part of Model; Just as its name implies, Controller is used to control. It receives the user's input and responds the user's requests by transferring Model and View. When users submit the form, Controller doesn't do any process or output, what it done is only to receive the diversified requests of all users. Then the system processes these requests by transferring Model groupware, such as browsing the userinfo, querying Q-A database etc. Finally it displays the data processed by Model through transferring View. The application of View in this system is the Servlet designed by relevant JSP .When users give the request, JSP will transfer the request to Servlet, and then Servlet could help to accomplish the functions such as raising a question or querying database by transferring the data processing through the relevant EJB groupware found by JNDI.

The following describes the system processing flow through the request of student raising a question. Firstly, the student should enter the JSP designed for login, after inputting the user name and password, JSP will transfer them to Servlet which is used to process the student's request of login. Then Servlet will help to lookup and validate its information to enter the page of student's module which includes the functions of querying or modifying the personal information, raising a question, scanning the forum and leaving word etc. Secondly, the student should choose the function of raising a question, the system will return the answers to JSP by transferring the Servlet in Q-A table according to the information of student's question. The detailed flow is described as following: when students submit the question to system by network randomly, the system should comprehend this question firstly, and then matches the question to the questions in Q-A database by computing the similarity between the question described in nature language and the questions in background Q-A database. If matching successfully, the correct answer or the correlative

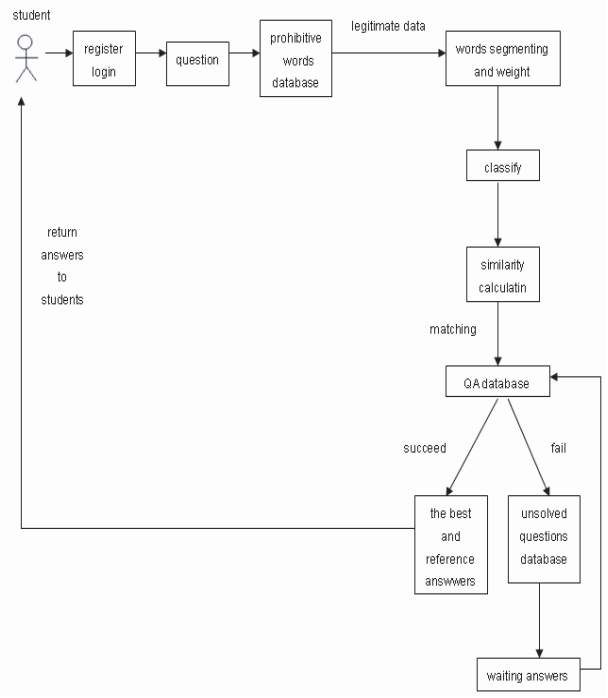

Figure 5. Flow chart of question answering system

answer will be returned to student, ending the question answering process; if matching unsuccessfully, it means there is no correlative answer in Q-A database, and then the question will be submitted to the database as an unsolved question, waiting to be answered by teacher later. Here, the system intercalates a prohibitive words database used to save the key words which is forbidden to be discussed in this system. The flow chart is as Figure 5.

The purpose of Presentation layer development is to provide users with dynamic Web pages when they visit the system, meanwhile it is in charge of communicating with EJB in Business, so that the user's request can be responded. EJB groupware is the architecture of J2EE platform [6]. Session Bean in J2EE application is applied to do some processes which are in server, for example, accessing the database, transferring other EJB module [4]. It is mainly used to communicate with fore-end Presentation layer, such as saving the user's login information, recording the data that user's operation to system (information such as writing the question into unsolved question database etc.), transferring Entity Bean to respond the user's request to Information layer. Entity Bean represents the data that preserved permanently, the typical type is the data saved in database. It can meet the user's demands by accessing those data, and it defines a series of data operation to fulfill all functions that system demand, such as raising a question, creating

or deleting users etc. The kind of entity is mainly determined by the condition of information table in background database, for example, this system includes the student information table, the teacher information table, Q-A table, words table and so on; Message-Driven Bean is just like an asynchronous information receiver who has realized some business logic, it is mainly used to process the asynchronous information. When JMS (Java information queue) receives a message, Message-Driven bean is transferred by EJB container. Message-Driven bean in this system is mainly in charge of sending important information to each function module entities, for example opening the individual information of students or teachers, amending the individual information and so on.

The last step is the design of Information layer. In the architecture of J2EE, system Information layer can be either database management system or other isomerous information system. It is mainly used for the memory management of enterprise information. It mainly includes the database system, the catalogue service and so on. The J2EE application groupware often need to access the EIS tier to obtain the required data information [3]. The data information of this system includes fundamental information of students, fundamental information of teachers, information of administrators, information of keywords, information of the Q-A, information of the unsolved questions and the fundamental information of forum etc.

In the system exploitation and deployment, WebLogic Server in BEA Company is used as the J2EE application server, which can help to implement all J2EE idiosyncrasies. WebLogic Server is an enterprise web application server, it can not only provide the application function of managing the J2EE application and other unattached application in the way of local mode and long-distance mode, but also can provide the application function of simplifying the structure of these applications. After structuring these applications, WebLogic Server also can provide engines for them.

5. Conclusion

The architecture of J2EE is core when designing and developing this system. It uses JSP, Servlets, EJB and JDBC database connectivity technologies. Compared with others, the most obvious advantage of this intelligent question answering system is that it can return accurate answers to users timely. In the other words, it can extract the optimal answer in the shortest time from Q-A database, it embodies the intelligence of extracting questions well. The main interface of the system is shown in figure 6.

Figure 6. Main interface of system

References

[1] U. S. A. An efficient algorithm for mining association rules. *Proceedings of the 21st International Conference on VLDB*, pages 432–444, September 1995.

[2] W. Z. Lei Ji, Li Li. *Master J2EE-Eclipse, Struts, Hibernate, Spring Conformity Application Cases*. POSTS and TELECOM PRESS, Beijing, 2006.

[3] Y. long Hao. *Technology of J2EE Programming*. QING HUA UNIVERSITY PRESS, Beijing, 2005.

[4] X.-l. Q. Qiang Zhao. *J2EE Application development(Weblogic+JBuilder)*. PUBLISHING HOUSE OF ELECTRONICS INDUSTRY, Beijing, 2003.

[5] I. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):11–12, January 2002.

[6] S. K. Vlada Matena. *Enchiridion of EJB Application-Development Based On Groupware*. QING HUA UNIVERSITY PRESS, Beijing, 2004.

[7] S. Wang. *Database technologies and On-Line analytical processing*. TECHNOLOGY PRESS, Beijing, 1998.