DESIGN OF AN INFORMATION INTELLIGENT SYSTEM
BASED ON WEB DATA MINING

Lect. RAMESH A. MEDAR and Prof. A. H. KULKARNI
Dept. of Computer science and Engg., Gogte Institute of Technology, Belgaum, India
Rameshcs.git@gmail.com

Dept. of Information science and Engg., Gogte Institute of Technology, Belgaum, India
ahkulkarni@gmail.com

ABSTRACT

To use the vast amounts of information efficiently on the Web to make the information processing intelligent, personalized and automatic is the most important applications of the current data mining technology. Model Driven Architecture (MDA) which is used for code generation has many benefits over traditional software development methods. In this paper, Web data mining process is introduced from the view of function, an intelligent mining system of information is built with combining the data mining. The concept of Web data mining is introduced where the role of MDA is defined. MDA using J2EE (Java to Enterprise Edition) to describe behavior of agents are used in this proposed architecture. Struts Framework provides a standard for developing MAS (multi-agent systems) and Web based applications.

KEY WORDS: Model driven, Intelligence, Struts

1. INTRODUCTION

With the rapid development of Internet, the information content on the Web has become very rich. A new technology is urgently needed to automatically find, extract and filter information from the Web resources. Web mining technology appeared. The purpose of Web Mining is to find and extract the potential useful model and the hidden information from the Web documents and Web activities. It combines together the traditional data mining technology and Web together, and can play a role in many ways, such as the mining of search engines, the development of search engines, improving and enhancing the quality and efficiency of search engines, determining authority pages, Web document classification, Web Log mining, intelligent enquiries and the establishment of Meta-Web data warehouse.

The WWW serves as a huge, wide, distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other services. With the rapid increasing of information in the WWW, the Web Mining has gradually become more and more important in Data Mining. Web Mining can be classified into three domains: Web Structure Mining, Web Content Mining and Web Usage Mining. There are generally three tasks in Web Usage Mining: pre-processing, knowledge discovery and pattern analysis.
2. WEB-BASED DATA MINING AND AGENT

2.1 Web-based data mining concept

Data mining is extracting potential, unknown, useful information, patterns and trends from abundant, incomplete, noise, fuzzy and random data which is used in practical application. Web mining comes from the development of data mining and it is a new field that the data mining technology used in Web information. It has new character compared with the traditional data mining. First, the objects of Web mining are a large number of Web documents which are heterogeneously distributed and each data source are heterogeneous; second, the Web document itself is semi-structured or unstructured and lack the semantics the machine can understand. The semi-structure refers to that although the data on the Web has certain structural, but has no specific model description, The data of each site is designed independently and the data itself has dynamic variability, consequently it is a type of non-total structured data. The objects of the traditional data mining are confined to be the structured data in the database and it used the storage structure such as forms of relationship to find knowledge, so some data mining technologies are not applicable to Web mining, even if useable it must be on the basis of Web documents pretreatment

2.2 Web data mining categories

In general, in accordance with the research object of Web mining, it can be divided into three categories which are Web content mining, Web structure mining and Web usage mining. As shown in Figure 1.

- **Web content mining** is the technology method the specific information users needed from Web information used to mine the Web content information. The latest research shows that two methods are mainly adopted and they are Information retrieve and database methods, in this paper a Web content information mining database model will be established

- The target of **Web structure mining** tends to the link structure of Web documents which reveals the personalized information contained in the document structure, and the type of data the Web structure mining processed is the structured data of Web. The structured data is the data to
describe the organization fashion of the Web page content, and the structure in pages can be
depicted as a tree structure by using HTML, in addition the structure between pages can be
depicted by using the hyperlink structure connected different pages.

(3) Web usage mining also known as Web log mining, its main objective is to find interesting
model from the Web visit record. And for the research in this area there are two main directions:
general access patterns track and personalized use record track.

2.3 Agent technology and multi-agent system

Agent refers to the entities which run in dynamic environment and have higher self-government
capacity. Agent software is a type of computer program which simulates human intelligence behavior
and provides the corresponding services. Agent has characteristics of autonomy, reaction, initiative and
sociality.

Multi-agent system is a multi-agent federation which is composed by a number of Agents and has
certain organization structure, as an effective method of solving complex systems, it uses parallel
distributed processing technology and modular design to divide the complex system into relatively
independent Agent subsystems and complete the solution of complex problems through the cooperation
between the Agents. The activities of the various Agent members are independent and autonomous, and
their own goals and behavior are not limited by other Agent members, they can coordinate and solve the
contradictions and conflicts of the goals and behavior among the Agent members through the means of
competition or consultations.

3. WEB-BASED DATA MINING PROCESS

The characteristics of Web data determine the immense challenge for effective data mining. According to the
characteristics of Web data and combining the general process of data mining (shown as figure 2), the Web data
mining process can be described as the five functional modules which are shown in figure 3, namely the data
acquisition, data preprocessing, data mining, analysis and evaluation and knowledge formulation modules. The
functions of each module are shown as follows.

3.1 Data acquisition

In accordance with the principles of themes relating, the data acquisition module selectively obtains
data from the outside Web environment to provide material and resources for the latter data mining. The
data source the Web environment provided includes the Web pages data, hyperlinks data and the data of
recording user visiting. According to different forms of data sources the Web data mining can be divided
into content-based mining, structure-based mining and the mining based on user usage. Each data
mining type may use different methods and techniques in the data acquisition process, but they have
same basic process. Generally, the data acquisition is composed by three relatively independent
processes which are data search, data selection and data collection.
3.2 Data preprocessing

Data preprocessing module mainly processes and reconstructs the source data the data acquisition received and builds the data warehouse of related theme to create basic platform for the latter data mining process. Data preprocessing is preparations for data mining and it mainly includes data scrubbing, data integration, data conversion, data reduction, etc.

3.2.1 Data scrubbing

The main role of data scrubbing is to remove the noise and unrelated data in the source data, process the omitted data and clean the dirty data, it also includes processing the repeating data and the data without value, and completing the conversion of some data types, such as converting the same kind of information from different sources into a unified storage.
3.2.2 Data integration

The main role of data integration is to handle the heterogeneous data from a number of sports environments and resolve the semantic ambiguity problem. Data integration is not a simple merger of the data, but a complex process of unified and standardized handling to the heterogeneous data.

3.2.3 Data conversion and data reduction

The main role of data conversion is to convert the data into the form which is suitable for data mining. And the main role of data reduction is to farthest reduce the data volume and enhance the efficiency of data mining algorithms through finding the useful feature of the data, and this is on the basis of fully understanding the mining tasks and the data content, under the premise of retaining the original data as possible.

3.3 Data mining

Data mining module is the core of the data mining system, and its main function is that using all kinds of data mining technologies to extract the knowledge model which is potential, effective and can be recognized by all people from the flood and pretreated data. Generally speaking, the ultimate goals of data mining only are description and prediction, the so-called description is that using comprehensible mode to express the attributes and characteristics information contained in the data; and the prediction is to find the discipline of the attributes according to their existing data value and then speculate a possible attribute value in the future. The data mining process is generally composed by three major phases: data preparation, mining operation, and interpretation of results. Data mining algorithms have certain on data, such as small redundancy, small correlation between the data attributes, small error ratio. But the data actually collected usually is chaotic, redundant and half-baked, so data mining must has the stage of data preparation for improving the quality of it. Mining operations include the choice of appropriate algorithms, operations of mining knowledge, last the operations of confirming knowledge; expression and interpretation stage is to analyze the results, extract the most valuable information. If the information obtained cannot make the decision –makers satisfied, the above data mining stage is needed to repeat.

3.4 Analysis and evaluation

Analysis and evaluation module is to analyze the credibility and effectiveness of the knowledge mode the data mining obtained, and educe evaluate conclusions to provide information support for the management decision-making of users.

3.5 Knowledge formulation

Knowledge expression module refers to the knowledge modes mined from the Web data by using data mining tools, and it will be shown with appropriate form to facilitate user acceptance and mutual exchange.
4. INFORMATION INTELLIGENCE MINING SYSTEM DESIGN

4.1 Information mining mode idea

Through the above analysis, two models for building are designed, the first is the mode of obtaining information through keywords (mode 1), the second is the mode of obtaining information through consecutive addresses (mode 2). The basic idea is: the first step, from the Web pages of particular Website analyzing and extracting some necessary parameters which are title or distinction sign of text beginning and ending, link address (including whether there is continuity, the base address corresponding to the relative address), and so on. The second step, according to these parameters obtaining the web page and extracting the required contents from it, this part of work is automatically completed by the designed system, and then the purposes of rapid and effective establishment of personalized information database will be achieved.

4.2 System structure

The model is designed aiming at the lack of the traditional search engine and its purpose is to design the information service intelligence system combining the above mode and idea on the basis of existing search engine. The system structure is shown in figure 4.

![Figure 4: Information intelligence mining system structure](image-url)
4.3 Subsystem functions

User interface Agent (interface Agent) is the interface of the users and computer for man-machine conversation and it can receive the user enquiries request, initiatively help users refine query request and display query results, in this process it synchronously learns and records users interest and habit. It also responsible for exporting the final results output to the corresponding users and displaying the content of local user database.

1) Problem analysis Agent. It analyzes user source and user preferences and submits the results to user local database, and synchronously it also explains the user needs according to the information of user local database.

2) Processing Agent. It receives problem analysis Agent and searches for the needing query results in the Internet according to the Boolean strings provided by the problem analysis Agent, and finally it will return the query result to information mining Agent.

3) Information mining Agent. It converts the basic information the processing Agent obtained into the directly understandable information through user interaction and intelligent processing (special mining algorithm), and then submits the information to the knowledge resources database.

5. IMPLEMENTATION DETAILS

The modules are designed both at the customer side and at the administrator side.

The administrator has the privilege to enter into the system by providing the username and password. The administrator can perform the data acquisition process, where he/she can enter the new url and the brief description of the url.. The added data will be present in the general data warehouse. In case of data preprocessing , the administrator separates the data automatically into the corresponding warehouses. The administrator can also view the content of different warehouses.(figure 5)

The client can view the system as a search engine. For example, if the client want to view about the web sites related to sports, then he/she have to enter the keyword ‘sports’ in the corresponding field and then click on the submit button. The search engine retrieves the information faster, because of the segregation of data into different warehouses.

6. CONCLUSION

As a new theme of data mining and data warehouse, Web-based is emerging research area. In this paper, the Web data mining process is introduced from the view of function, an intelligent mining system of network information is built with data mining technology and the functions of the various subsystems of it are also introduced.

For the Web-based data warehouse and data mining technology, the development of the Internet not only presents challenge but also provides a broad application prospects. And with the continuous development of the computer, network and communications technology, the research of Web-based data
mining will be further in-depth, and it also will play an increasingly important role in Web site design, personalized e-commerce services, etc.

![Diagram](image)

**Figure 5 : Administrator modules**

7. RESULTS

![Image](image)

**Figure 6 : Login page**
Design of an Information Intelligent System based on Web Data Mining

Figure 7: Data addition

Figure 8: General data warehouse
Figure 9: Content of General data warehouse

Figure 9: Data separation
Design of an Information Intelligent System based on Web Data Mining

Figure 10: Sports data warehouse

Figure 10: Content of Sports data warehouse
Figure 11: Client entering the keyword

Figure 12: Search results
8. REFERENCES


