Research on Data Acquisition and Storage Management System Based on Web Mining Technology

Jiale Zhang

School of Software, Zhengzhou University, 450002

Keywords: Web mining; Data collection; memory management

Abstract: Data acquisition and storage management is an important part of information system. With the wide popularization of information technology tools in various departments, the data acquisition mode has changed. From the perspective of user behavior, this paper discusses the problems faced by the traditional Web usage mining research based on Web logs, and makes an in-depth analysis of the data collection technology of Web usage mining based on user behavior. The Web database is implemented by ASP technology, and the remote intelligent terminal is implemented by single chip microcomputer system. Users can monitor the status of field devices through browsers. This paper focuses on the active server-side data acquisition method and client-side data acquisition method.

1. Introduction

With the maturity of network technology and information technology, Ethernet has been widely used in many fields such as data acquisition and transmission, data communication, etc. Network management of data acquisition and remote monitoring technology have become the inevitable trend of industrial development [1]. M2M network communication technology has become a research hotspot at home and abroad, and Web technology provides a solution for industrial equipment to realize network management. The early Web usage record mining mainly uses the information left in the Web log when users access the Web to obtain the user browsing path. This method is more suitable for static pages and cannot be well applied to the current Internet dynamic applications [2]. At the same time, with the development of communication network and the increasing number of unattended computer rooms, people have put forward higher requirements for the normal operation of power equipment and computer room environment monitoring system. The characteristics of the above systems are that the sites are scattered and the normal operation of the sites is extremely important [3]. Web uses data to mine hidden and interesting patterns. Web usage mining process is similar to data mining process, which is divided into four stages: data collection, data preprocessing, pattern discovery and pattern classification. This paper analyzes the design of data collection and online publishing system based on Web, and analyzes and introduces some research topics related to this system.

2. Questions Raised

Due to the lack of precise collection mechanism and method of user usage data in Web usage mining, currently, the common research methods on Web usage mining are based on Web log files, that is, Web log files are taken as data collection objects [4]. In traditional Web usage record mining, the data set used in mining is user access log, which exists in server, proxy server and client respectively. Web service is an application program programming interface that an application program can call externally. The application program that calls this Web service is called the client, while the application program that provides this Web service is called the server [5]. On the surface, a user's click on the browser will send a request to the server, calling a page of the website, but at the same time it will open all the objects involved in the page, and each object will generate a record in the log file. In the past, when the computer integration technology was not developed, we used

manual on-site control and maintenance, which not only wasted manpower and material resources, but also had low efficiency. Therefore, based on the existing communication network, the monitoring system was used to construct remote monitoring with low cost, strong reliability and easy operation and configuration through different site resources. It is not advisable to simply collect Web usage data directly from the log file of the Web server for Web usage mining research, which has some drawbacks:

3. Massive logs are massive garbage data for Web users to use data.

Through the research on the recording methods of Web log files, it has been known that a user's request for a page (i.e. one click) on a Web server often generates multiple or even hundreds of records in the server's log files. The server API is an extended CGI tool. User applications written with the API are compiled into dynamic link library DLL, and the Web server runs it in thread mode, thus saving the communication overhead between processes [6]. This is mainly because user access is not one-to-one with server resources, and the log of the Web server may record the situation where a user submits requests in multiple clients, or the situation where multiple users submit requests in one client. The usage mining based on Web log must preprocess these massive data through data purification operation, and the remaining available data is about 5% ~ 10%. When searching for services, the Agent makes a query request to UDDI system in SOAP message format and generates a query Agent to receive SOAP messages required by Web services.

3.1 Incomplete page browsing record.

The Web log file is a record of every HTTP request that arrives at the Web server. There are various caching mechanisms on the Internet, which lead to the incompleteness of user page browsing records in the Web log files. Since the Web server processes the requests of multiple users concurrently, it is difficult to identify all users' access sessions, which is generally based on the assumption of user browsing behavior [7]. In some places, the cache can provide 20% ~ 50% of the request service, which means that the server log loses a lot of user browsing records. Web services shield the differences between heterogeneous systems, and external collection requests from Agent that are responded to through service interfaces. UDDI is a directory service that enterprises can use to register and search Web services. Session identification only distinguishes the session periods when users visit the server one by one, but it does not include all the pages that users have visited. This is because the client has a cache. Therefore, it is necessary to infer and perfect the session periods when users visit the server, that is, path completion.

3.2 Unable to get accurate page browsing time.

Calculating the user's stay time on a page based on the Web log is generally obtained by subtracting the current HTTP request's time domain from the next HTTP request's time domain. However, due to network transmission and server response time, this method is not accurate in calculating browsing time. The disadvantage of server-specific API is that they are not compatible with each other. It is more difficult to develop API programs than CGI programs and to debug them. JDBC is a general and low-level API that supports basic SQL functions. The key technology is a set of Java interfaces implemented by drivers [8]. Statistical analysis: session files can be analyzed to obtain statistical information on web browsing, browsing time, path length, etc. Association rule mining can discover web pages frequently visited together in a conversation, and provide decision support for marketing or reorganizing website content; When the user looks at the webpage in the cache, the time spent browsing the webpage in the cache will also be accumulated to the page returned by the previous HTTP request; When the user leaves the website, the browser does not make an HTTP request, resulting in the viewing time of the last page being unavailable, etc.

4. Research on Web Mining Technology and Storage Management System Technology

4.1 Classification of data collection.

W3C organization defines relevant concepts for describing users' usage behavior on the Web, including users, user sessions, page files, page views, server sessions, scenarios, etc. Pattern evaluation is used to transform the discovered patterns into useful knowledge. Generally, meaningless or worthless patterns are removed in combination with domain knowledge. On the one hand, the collected data are sent to STM32F103 server and displayed on the LCD screen in real time; on the other hand, the remote client's Web browser is used to input the IP address corresponding to the microprocessor to access the data, thus realizing the function of multi-point collection through networking [9]. These concepts are the basis for understanding the characteristics of Web usage data. The user's access behavior results in a lot of data containing behavior information, which interacts in the Web environment. With the help of a common Web browser, the system runs on a software platform. The system appears in the form of an Applet embedded in an HTML file. Communication with XML file, regular data collection can be carried out at any level [10]. The register () method is used to register the web addresses of departments at different levels. The send () method is used to send an XML file to register an account; The get () method is used to obtain data. The main research direction of data acquisition technology is how to effectively obtain usage data from Web interaction environment. Figure 1 shows a brief description of the Web interaction environment.

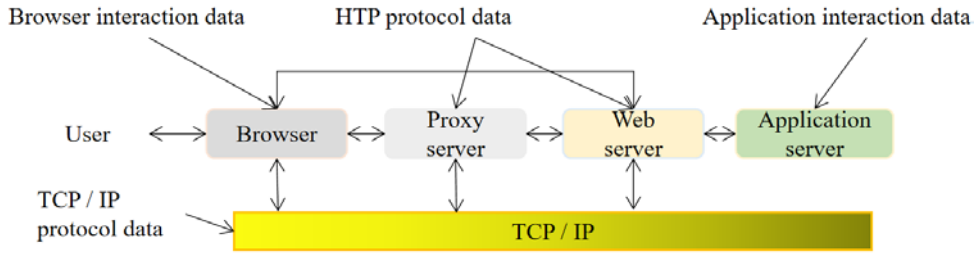

Figure 1 Web interaction environment

As can be seen from Figure 1, data mined by Web usage can be collected from client browsers, proxy servers, Web servers, application servers and even TCP/IP network layers. The Web pages provided by the monitoring center are respectively user authentication login page, identity authentication and function selection interface, and site equipment parameter real-time monitoring interface. Can do user clustering to further provide personalized services for users, can also do web clustering to assist in the establishment of web groups with related content; Sequential pattern analysis can discover the temporal correlation between web pages within a session and help predict the access of web pages. Browser interaction data includes events occurring on the browser interface and contains rich user behavior information, which is collected at the client. When a user creates a new report file or opens an existing report file, the system will construct an instance of the SC Report Data object. Each table object in the periodic report appears as an element variable of the SC Report Data object. HTTP protocol data is data generated by communication between browser and Web server, including requested URL and relevant protocol data. No matter what the platform and programming language are, there is no need to reconfigure the operating system, database system and application software, allowing access to files in different private networks and free communication with HTTP protocol.

4.2 Active server data acquisition and storage management.

The so-called passive data acquisition refers to a method that does not actively adopt appropriate methods to acquire corresponding data according to the target of data mining to be carried out, but simply extracts the required source data from the existing data. The system performs authentication. First, the system searches the table manager for the user name and password. If the user is a legal user and the password matches, the system enters the contents of the selection page. The user access

record is collected by the server application. User access records organically combine the contents of Web logs with shopping records and inquiry records in e-commerce, which can effectively reflect the user's interest access mode. Then send a request to the Web service in the required format to transfer relevant data. In the process of completing data declaration through the data declaration system, users often need to check the data of currently edited files, which not only ensures the integrity of user data, but also ensures the correctness of data. A series of log files automatically recorded by Web server and proxy server software are ready-made data sources, and they also contain most user usage data, such as user's IP address, URL accessed, date and time of access, access method, etc. Therefore, an ASP file named check.asp is used for identity token keeping authentication, and then this file is executed at the beginning of each Web page for authentication.

The database correspondence setting function is used to set the correspondence between the data dictionary and the storage of the table structure in the database, which provides great flexibility for the data warehousing function of the system, establishes the correspondence between the table items on the interface provided to the user and the table items stored in the database, and allows the database manager to modify for different report types, thus ensuring the reusability of the system well. The flow of data warehousing can be simply represented by Figure 2 .

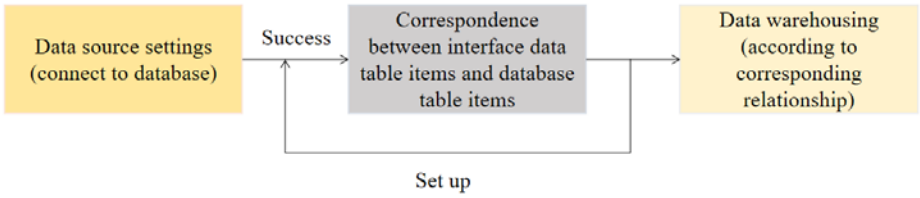

Figure 2 Database warehousing process

In the previous discussion, the disadvantages of simply collecting Web usage data directly from Web server log files for Web usage mining research have been clarified. One of the methods to solve these problems is to adopt active server-side data collection. Because the amount of data collected is usually large, it is necessary to provide corresponding controls for the data in the graphical user interface. Users can set the data values by operating these controls. The Web server provides Cookie Tracking function. When the user browses the server, the server will automatically save a very small text file, i.e. Cookie file, on the hard disk of the user terminal. The equipment parameter data of each site collected from the intelligent terminal are put into the database param table, scripts are written in ASP, these parameters are called from the database, put on the Web page, and provided to users. However, for the first visit, the visit of unregistered users and the visit of registered users without login, the processing is relatively rough and there are deficiencies. In addition, the collection of "user access records" by the server data collection program increases the load on the server, and the logs of the Web server are not utilized. In the process, it is necessary to switch between different panels to complete the input of different forms of data for periodic reports, while only one panel can be active at a time in the palette. Therefore, the design idea of component library is considered and applied to the design of graphical user interface.

4.3 Client data acquisition and storage management.

Client-side data collection is more advantageous than server-side data collection because it is based on the user's behavior source and can capture the user's behavior comprehensively and accurately. In Web remote systems, automatic refresh of web pages is used to complete automatic update of data. The key to automatic web page refresh is to determine the web page address that needs to be refreshed. Data warehouse is a topic-oriented, integrated, relatively stable data set that reflects historical changes. It is mainly used to support decision-making. However, when the system is running, especially in the panel switching process inside the palette, there will be a certain delay, and this delay process is very short. The measurement of the user's browsing path and browsing time can be very accurate, avoiding the complicated process of user identification, session identification, path supplement and the like, which affects the accuracy. Therefore, a power-down protection circuit is

adopted in the design of the intelligent terminal. Important data are stored in EEPROM in time. After power-on reset, important parameters are read into RAM through a program. The aim is to change the content of the web page from "website" to "user", and adjust it automatically as much as possible to cater to the browsing interest and purchasing mode of each user, so as to facilitate users (visitors to the website) and promote purchasing volume.

After introducing the active data acquisition method oriented to user behavior, the process can be divided into four stages, namely data acquisition, data processing, pattern discovery and pattern analysis. The process model of Web usage mining is shown in Figure 3.

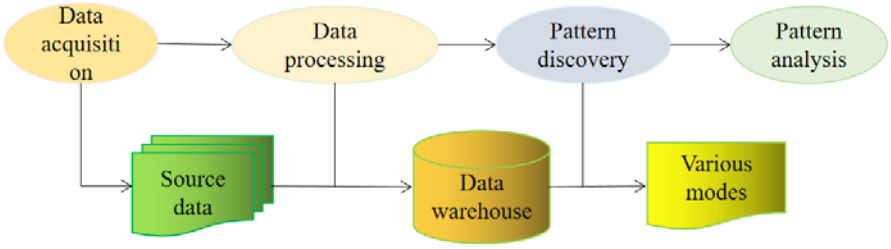

Figure 3 Web usage mining process model

When using Java Applet technology to collect user's usage information at the client, the client browser needs to download and install Sun JVM plug-in and add applet program code at the beginning of each web page to be tracked. In parallel interface mode, P0 port of MCU is directly connected with data ports D0-D7 of EMODEM, while P0 port of MCU is connected with address line 2 of EMODEM through address latch. When a user accesses a Web server, the Applet program will be downloaded to the client and run on the browser. The program is programmed using the powerful Java class library, which is relatively easy to implement in data collection and data return. The contents of the user access record include the user identification, the user's IP address, the usage agent, the access date and time, the source page of the page request, the object URL, the keywords searched, the type of access object, the user's actions (viewing, purchasing, giving up purchasing, etc.) and the commodity number. After the database correspondence is completely designed, the data can be put into storage. The data warehousing operation is also based on the correspondence table set by the administrator. The premise of applying Plug-in technology is that the browser must install Plug-in plug-in in advance, otherwise it will cause the acquisition program to be unable to run and the user usage information will not be available.

5. Summary

The research on Web usage mining is still a new research field. How to collect user usage data accurately, timely and comprehensively is an important premise and foundation for the research on Web usage mining. The design scheme proposed in this paper is a kind of practice to realize data collection and storage management on the Web. In the design, a preliminary design idea has been preliminarily realized through relevant technologies. This method has been used in some specialties. Starting from the overall Web interaction environment, this paper systematically expounds the data collection methods and classification, and makes a detailed analysis on the data collection methods and implementation technologies of active server and client. Web usage record mining has a wide range of application fields, of which e-commerce is one of the most important fields. Web mining application system oriented to e-commerce has become a research hotspot, and we will continue to conduct more in-depth research.

References

[1] Kebin J, Hanjing L I, Ye Y. Application of Data Mining in Mobile Health System Based on Apriori Algorithm[J]. Journal of Beijing University of Technology, 2017, 43(3):394-401.

[2] Kammoun M A, Rezg N. Toward the optimal selective maintenance for multi-component systems

using observed failure: applied to the FMS study case[J]. International Journal of Advanced Manufacturing Technology, 2018, 96(2):1-15.

[3] Liu X. The Application of Data Mining Technology in the Teaching Evaluation in Colleges and Universities[J]. Journal of Computational and Theoretical Nanoscience, 2017, 14.

[4] Sakurai Y, Matsubara Y, Faloutsos C. Mining and Forecasting of Big Time-series Data[J]. Japanese Journal of Psychonomic Science, 2017, 35(S1):919-922.

[5] Liu Y, Weng X, Wan J, et al. Exploring Data Validity in Transportation Systems for Smart Cities[J]. IEEE Communications Magazine, 2017, 55(5):26-33.

[6] SM Yim. Web-Based Collaborative Writing in L2 Contexts: Methodological Insights from Text Mining.[J]. Language Learning & Technology, 2017, 21(1):146-165.

[7] Liu D, Cai S, Guo X. Incremental sequential pattern mining algorithms of Web site access in grid structure database[J]. Neural Computing and Applications, 2017, 28(3):575-583.

[8] Jin J C, Zhang J, Lv Z C. A novel gradient climbing control for seeking the best communication point for data collection from a seabed platform using a single unmanned surface vehicle[J]. Frontiers of Information Technology & Electronic Engineering, 2019, 20(6):751-759.

[9] Luyao F, Jing Z, Xiaolong C, et al. Open source big data framework in marine information processing[J]. Science & Technology Review, 2017, 35(20):126-133.

[10] Huang Wei. Research and design of school data acquisition platform based on Web service [J]. Wireless Internet Technology, 2017(19):93-95.