The Study on the Preprocessing in Web Log Mining

Ma Shu-yue, Liu Wen-cai
College of Information
Jiu Jiang University
Jiu Jiang, Jiangxi Province, China

Wang Shuo
College of Information Engineering
Inner Mongolia University of Technology
Inner Mongolia, China

Abstract—**According to the Web log mining, the site administrators can control the network traffic and understand the user access modes. Then they can further improve the performance of Web systems and optimize the system design of Web sites by using these information. However, the Web log data doesn't perform the data mining directly in most cases because of the messy and redundant content and other reasons. This paper analyzes the data pre-processing on Web log in order to meet the needs of data mining. At the same time, it also puts forward some reasonable processing means.**

Keywords-Web log mining; data cleaning; user identification; transaction identification segmentation

I. Introduction

The development of Internet has promoted the progress of WWW. Following the successful application of the data mining techniques in the traditional database fields, people have begun to study the Web-based data mining technology (Abbr. Web Mining). The Web log mining is a technology which applies the data mining technologies to the log files in Web server in order to find the browsing patterns of users and analyze the site usage. And it can be also used to assist the site managers to optimize the sites.

The Web server logs record the user's information of accessing the site. The typical Web server logs contain the following information: IP address 、request time 、 method(eg GET) 、URL of the requested files 、HTTP version 、 return codes 、 the number of bytes transferred 、the Referrer's URL and agents. However, the data in Web logs isn't precise because of the existence of local cache, proxy servers and firewalls. So it is difficult to make a mining directly on it and we may get some wrong results.

II. Web Log Data Preprocessing

The Web logs are often available in two formats: CLF (Common Log Format) and ECLF (Extended Common Log Format). The most basic data fields of ECLF format data contain the client IP, User ，Time ，Request ，Status ，BytesRecvd ，BytesSent ，Process Time ，Reference ，Agent. Among them, the User has the data only when the request files need to be certificated. Time records the time of issuing files that the server responses to the user request. Request records the method of user request, URL and the used protocol. Status is recorded by the server which shows the response to a request. BytesRecvd records the number of bytes that the users send to the server when they make a request. BytesSent records the number of bytes that the server which processes the request has sent. Reference records the URL which has sent the requests, and when the users enter an address or utilize the bookmark to access it, the reference is empty. Finally, Agent records the operating system and the browser type of users.

In order to get the suitable Web log data to perform the data mining, we must undertake a series of operations on the original Web log files such as the log consolidation and data cleaning, user and transaction identification, data integration and so on. The basic process of Web log mining preprocessing is shown below:

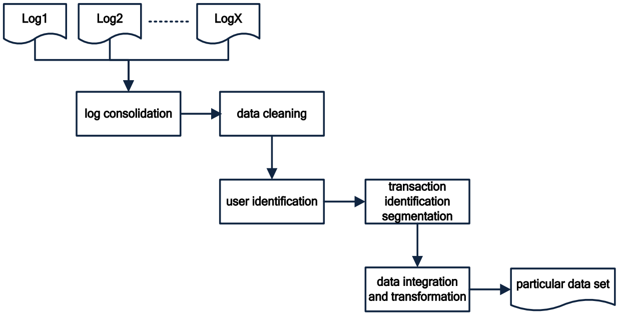

Fig 1 Web Log Mining Preprocessing

III. Log Consolidation and Data Cleaning

A. Log Consolidation

In the previous studies, the consolidation of Web logs is relatively simple. This is because under the circumstance that the Web service content is limited, as for a single Web server, the log files are just generated according to some certain naming rules and the different time. The log consolidation is just a simple timing accumulation of these files.

With the enrichment and expansion of Web application content, many large-scale Web services include more and more contents. Then most of Web services have employed the automated multi-server load balancing architecture. At this time, the logs which are served by the same Web are usually scattered and stored in different servers. In light of these circumstances, we need to periodically synchronize to the background or a special log server through the certain means. After the consolidation, we should use the log analysis tools to analyze.

B. Data Cleaning

Data cleaning is also known as data cleanup. Data cleaning is a process in which it will eliminate the data errors、 inconsistencies and solve the object recognition problems. Data cleaning contains the null value、noise and data processing、the inconsistent data processing and some others. The inconsistencies of data lead to the reduction of credibility of the data mining results. The data cleaning removes the noise or irrelevant data, and also processes the missing data field in the data.

Being mainly against the problems of the irregularity of data in multiple data sources、ambiguity、 duplication、non-integrity and some others, the data cleaning accordingly performs the cleaning operations for the error data. Data cleaning can improve the quality of data, thus enhance the accuracy and performance of the subsequent data mining process. Because the high-quality analysis and decision must depend on the good-quality data, data cleaning is an important step in the data mining.

1) Page Element Cleaning: In most cases, only the HTML files in logs have associated with the user session. The users don't explicitly request the graphics files on the pages and they are usually automatically downloaded according to the hypertext reference tags of HTML. Because the purpose of Web log mining is to obtain the behavior patterns of the user and it doesn't care the files which the users don't explicitly request. So it is necessary to remove the irrelevant data by checking the suffix of URL. For example: the files in the log which have the suffix gif、jpeg、jpg and map should be removed.

However, as for some special image sites and news pages, the information may be the focusing elements which the users are of great interest. So the action of simply deleting these records hasn't satisfied the data mining's further requirement for granularity. Therefore, the in-depth study has been made for the judgment and further processing of the image information.

On account of the image files which have the suffix jpg and peg, because the size of the pictures can be visually shown by the size of images in this picture compression standard and at the same time, we can also get the file size in log records. Therefore, it is doable to set the lower threshold of the image file's size when the log files are cleaned. The files which are higher than this value can be considered as the valid page request objects and can be retained, and the files should be removed when they are below this value.

On the other hand, the files with suffix gif include a great number of animated images. These images occupy a larger space but their sizes are small. So as for the gif files, we should directly remove them or take a method of analyzing the image's size instead of judging them based on the file's size.

At the same time, the script files with the suffix cgi、js and js should be deleted because they have no practical significance to the analysis and processing of data mining.

2) Cleaning of Other Information: There are still existing a great number of classes(*.css，*.xsl，*.xsd，*.dll, etc)、the systematic dynamic link library and some others in Web logs which are generated by system calls. These logs will appear when the Web server system starts up and backups. They should be also cleared because these logs have nothing to do with our analysis of the customer's usage of the sites.

IV. User and Transaction Identification

A. Determination and Identification of the Users

The determination of user in Web logs is a difficult point. Generally speaking, it is difficult to accurately determine a user unless we employ some special techniques. There are both static pages and dynamic pages existing in Web services. In the process of site operation, it will be simple to determine the users if the SessionID is embedded and the users are required to register or the Cookie is written to the client. However, if they are static pages, the only thing we can do is just to use the IP address to analyze the user's access on logs. The agent types and some temporary information are combined together to identify a user and determine the users. Firstly, if the IP addresses are identical, but the browser software or the operating systems are different in Agent information, then two different users can be assumed. Secondly, if IP address and the Agent information are identical, then we should judge whether there is a connection between the pages which are requested to access and the pages which have been accessed. If there is no direct link between them, then we can assume that there exist multiple users in the machine which accesses the Web sites.

The reality is that the professional service providers usually set up a proxy server to increase the utilization of IP address because the IP address is limited, but they have provided a great number of Internet users and the time of surfing the Internet is uncertain. However, it is different in Education Networks. According to the incomplete statistics, there are more than 90% of Internet users using the fixed IP in Education Networks and only a few people access the Internet by sharing the proxy server. Therefore, we can make the user's access to IP as a symbol to distinguish the users, assuming a user IP for every user.

B. Transaction Identification Segmentation

Generally speaking, the task of transaction identification is to break a large transaction down into several smaller ones or combine the small transactions into a large one. So the main methods of transaction identification contain segmentation and consolidation. In the Web log mining, the user session is the only object with the characteristics of natural services. However, the granularity of it is too coarse for the mining association rules and some other methods. It is necessary to use the segmentation algorithm to translate them into smaller transactions.

At present, there are three transaction identification segmentation algorithms: Reference Length、Maximal Forward Path and Time Window. The first two algorithms are to identify the semantically meaningful transactions and the last one is mainly used as the benchmark for comparing the other two algorithms.

1) Reference Length: Reference Length refers to the time that the user browses the page, and it can be regarded as the time interval between the current page request and the next request without considering the case of network delay. Reference Length transaction identification algorithm assumes that the time the user spends on a page is in connection with the situation that the page is a content one or a navigation one for the user. The qualitative analysis of page reference length in server logs shows the exponential distribution. Assuming that we have known the percentage of navigation pages in logs, and the boundaries between the navigation pages and the content pages can be obtained by calculating. Then we classify the page reference of each user session to get the content pages, which means we have got the corresponding transaction of the user session.

2) Maximal Forward Reference: At times, some pages contain more hyperlinks which are the information the users are of great interest, so they are usually considered as the content pages. At this moment, the maximal forward reference path is employed to define the transactions. For every user session, taking the beginning page as a starting point, every maximal forward reference path is also regarded as a transaction. From the first reference to the backtracking at somewhere, here every transaction is defined as a visit of a set of pages. The forward guideline is defined as the page which never appears in a collection of transactions and the backward guideline refers to the page which has emerged in the pervious transaction. When a forward guideline appears, a new transaction is started. This method defaults that the maximum forward guideline page is the content page and the page which guides the maximum forward guideline page is regarded as the navigation one.

V. Data Integration and Transformation

As for the access sequences of the user which are obtained by data cleaning, they may go beyond a long time period. So it is possible that the user has visited the site more than once in this period. The purpose of transaction identification is to divide all the visit sequences of user into multiple separate user visit sequence. The simplest method to obtain this division is to define a time period. If the access time interval that the user requests any two adjacent pages goes beyond this time period, then it will consider that the user has started a new session. And in general, this time period is selected as 30 minutes.

At the same time, the attribute value of the same real world entity from different data sources may be different. This difference may be due to the different representation, ratio or coding. For example, the representation of time is inconsistent in the server logs and agent logs or the error referenced logs. Even the representation of the same page is also inconsistent in the server logs. At this time, we should transform them into the same representation in order to improve the accuracy and speed of the subsequent mining. Data transformation is to convert the data into the form which is suitable for mining. The common methods contain smooth, aggregation, standardization, data generalization, attribute constructors and some others. After the data transformation, the corresponding data mining could be performed for the formative data, such as: association rules mining, sequential pattern mining and so on.

VI. Conclusion

In the field of Web log mining, the domestic and international researchers have done considerable research and practice. However, the rapid development of the Internet makes these studies lag behind the current situation. In terms of the data preprocessing in Web log mining, there also exist such problems. Many problems have been highlighted with the development. And a single data granularity will be appeared when we use the existing methods so that it can't meet the requirements of data mining for the further refinement. At the same time, the embedded animations in Web pages and other page elements which meet the new standards can be combined with the logs to become the concerns. Therefore, the preprocessing before the data mining in Web logs should become a more important research.

References

[1] Zhuang Like, Kou Zhongbao, Zhang Changshui. Session identification based on time intervals in Web log mining [J]. Journal of Tsinghua University (Science and Technology), 2005，45 (1) : 115 - 118.

[2] Bao Yu, Huang Guoxing, Zhang Zhao. Mining Web Logs to Improve Website Organization [J]. Computer Engineering, 2003，7.

[3] Zhang Xiaodi. A Revised Data-Preparation Method for Web Usage Mining [J]. Computer Engineering and Application, 2006，17:160—162.

[4] Brigitte Trousse. AxIS Project[EB /OL]. http: / /www2sop. inria. fr/ axis/ axislogminer / ， 2003 - 12- 10 /2005 - 08 - 25.

[5] Dell Zhang，A novel Web usage mining approach for search engines，Computer Networks， 2002， 39:303-310