Research and Development of Data Preprocessing in Web Usage Mining

Li Chaofeng
School of Management, South-Central University for Nationalities, Wuhan 430074, P.R. China

Abstract Web Usage Mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. There are several preprocessing tasks that must be performed prior to applying data mining algorithms to the data collected from server logs. Data preprocessing is the process to convert the raw data into the data abstraction necessary for the further applying the data mining algorithm. This paper presents several data preparation techniques that can be used to improve the performance of data preprocessing in order to identify unique users and user sessions. These techniques and algorithms have been proved valid and efficient by experiments. Finally, we conclude this paper and propose the future research directions.

Keywords Web Usage Mining; Web log; Data Preprocessing; User Session

1 Introduction

Web Usage Mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications[1]. The results of Web Usage Mining can be used in personalization, system improvement, site modification, business intelligence, usage characterization and so forth. Generally, Web Usage Mining consists of three processes: data preprocessing, patterns discovery and patterns analysis. As the data sources of patterns discovery, the results’ quality of data preprocessing influences the results of patterns discovery directly. Better data sources can not only discover high quality patterns but also improve the algorithm of Web Usage Mining. So, data preprocessing is particularly important for the whole Web Usage Mining processes and the key of the Web Usage Mining’s quality[2]. However, Each type of data collection used in data preprocessing differs not only in the terms of the location of the data source, but also the kinds of data available, the segment of population from which the data are collected, and it’s method of implementation.

The research on data preprocessing of Web Usage Mining is a focus field nowadays. This paper attempts to present the process of data preprocess in data preprocessing of Web Usage Mining. The organization of our paper is as follows. In Section 2, we review the related researches in Web Usage Mining. We analyze the processes of data preprocessing in Web Usage Mining in detail and propose the algorithms of each step of data preprocessing in Section 3. In Section 4, an experiment is given to verify the effectiveness and efficiency of our algorithms. Finally, Section 5 concludes this paper with suggestions for future research.

2 Related Works

More and more researchers focus on Web Usage Mining recent years[1,3,4,5]. However, data preprocessing in Web Usage Mining has received less attention than its importance warrants. Robert Cooley, Bamshad Mobasher and Jaidep Srivastava presented methods for user identification, session identification, page view identification, path completion, and episode identification[1]. They proposed some heuristics to deal with the difficulties during data preprocessing. Bettina Berendt and her colleagues compared time-based and referrer-based heuristics for visit reconstruction[6]. They found that a heuristic’s appropriateness depends on the Web site’s design and on the visits’ length. Doru Tanasa and Brigitte Trousseau proposed advanced data preprocessing for intersites[7]. They offered the possibility of jointly analyzing multiple Web server logs. However, none of these methods are without serious drawbacks.
3 Data Preprocessing in Web Usage Mining

Ideally, the input for the Web Usage Mining process is a user session file that gives an exact account of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is the set of the page accesses that occur during a single visit to a Web site. However, because of the reasons we will discuss in the following, the information contained in a raw Web server log does not reliably represent a user session file before data preprocessing. Generally, data preprocessing consists of data cleaning, user identification, session identification and path completion, as shown in Figure 1.

![Fig. 1 Phases of Data Preprocessing in Web Usage Mining](image)

3.1 Data Cleaning

This task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data needed to clean: accessorial resources embedded in HTML file, robots’ requests and error requests.

1) Accessorial Resources. Because HTTP protocol is connectionless, a user’s request to view a particular page often results in several log entries since graphics and scripts are downloaded in addition to the HTML file. Since the main intent of Web Usage Mining is to get a picture of the user’s behavior, it does not make sense to include file requests that the user did not explicitly request. Elimination of the items deemed irrelevant can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG, css and map can be removed. In addition, common scripts such as the files requested with the suffixes of “.cgi” can also be removed.

2) Robots’ requests. Web robots (also called spiders) are software tools that scan a Web site to extract its content. Spiders automatically follow all the hyperlinks from a Web page. Search engines such as Google periodically use spiders to grab all the pages from a Web site to update their search indexes. To remove robots’ request, we can look for all hosts that have requested the page “robots.txt”.

3) Error’s requests. Error’s requests are useless for mining process. They can be removed by checking the status of request. For example, if the status is 404, it is shown that the requested resource is not existence. This log entry in log can be removed then.
3.2 User Identification

A user is defined as the principal using a client to interactively retrieve and render resources or resource manifestations. User identification is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, it’s difficult because of security and privacy. In our experiment, we use the following heuristics to identify the user:

1) Each IP address represents one user;
2) For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, an IP address represents a different user;
3) Using the access log in conjunction with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, there is another user with the same IP address.

3.3 User Session Identification

A user session means a delimited set of user clicks (click stream) across one or more Web servers. The goal of session identification is to divide the page accesses of each user into individual sessions. At present, the methods to identify user session include timeout mechanism and maximal forward reference mainly. The following is the rules we use to identify user session in our experiment:

1) If there is a new user, there is a new session;
2) In one user session, if the refer page is null, there is a new session;
3) If the time between page requests exceeds a certain limit (30 or 25.5 minutes), it is assumed that the user is starting a new session.

3.4 Path Completion

As the existence of local cache and proxy server, there are many important accesses that are not recorded in the access log. The task of path completion is to fill in these missing page references. Methods similar to those used for user identification can be used for path completion. If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. If the page is in the user's recent request history, the assumption is that the user backtracked with the “back” button available on most browsers, calling up cached versions of the pages until a new page was requested. If the referrer log is not clear, the site topology can be used to the same effect. If more than one page in the user's history contains a link to the requested page, it is assumed that the page closest to the previously requested page is the source of the new request.

4 Experiment

To validate the effectiveness and efficiency of our methodology mentioned above, we have made an experiment with the web server log of the library of South-Central University for Nationalities. The initial data source of our experiment is from May 28, 2006 to June 3, 2006, which size is 129MB. Our experiments were performed on a 2.8GHz Pentium IV CPU, 512MB of main memory, Windows 2000 professional, SQL Server 2000 and JDK 1.5. Table 1 is the results of our experiment.

As shown in Table 1, after data cleaning, the number of requests declined from 747890 to 112783. Figure 2 shows the detail changes in data cleaning.

<table>
<thead>
<tr>
<th>Table 1. The Processes and Results of Data Preprocessing in Web Usage Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entries in raw web log</td>
</tr>
<tr>
<td>747890</td>
</tr>
</tbody>
</table>
In Figure 2, Bar chart 1 represents the initial requests in raw web log. From Bar chart 2 to 6 represent the requests after removing the log entries with filename suffix “gif” or “GIF”, the log entries with filename suffix “jpg” or “jpeg”, the log entries with filename suffix “css”, robots’ requests and error’s requests.

![Fig. 2 Processes of Data Cleaning](image)

Figure 3 is the processes of user identification. Bar chart 1 is the number of users identified only by IP addresses. Bar chart 2 is the number of users with the same IP address and agent. Bar chart 3 is the number of users considering local cache and proxy server. The values of Bar chart 1, 2 and 3 are 4575, 5440 and 55052 respective. The results of user identification accord with the facts because we set many proxy servers actually in our university.

![Fig. 3 Processes of User Identification](image)

Finally, on the basis of user identification’s results, we have identified 57245 sessions by a threshold of 30 minutes and path completion.

5 Conclusion

This paper has presented the details of data preprocessing tasks that are necessary for performing Web Usage Mining, the application of data mining and knowledge discovery techniques to WWW server access logs. We give some rules in every phase of data preprocessing in order to design and implement them easily. Our experiments have let us estimate data preprocessing’s importance and our methodology’s effectiveness. It not only reduces the log file size but also increases the quality of the available data. However, many problems remain such as data collection, applications of some heuristics in some phase of data preprocessing, the accuracy of user identification and session identification, applying the results of data preprocessing to patterns discovery and so on. We’ll focus on solving these issues in the future.
References


