Analysis of Preprocessing Methods for Web Usage Data

Liu Kewen
School of Computer and Informaton Engineering
Harbin University of Commerce
Harbin, China
liukewen@hrbcu.edu.cn

Abstract—**Web usage mining is to analysis Web log files and discover user accessing patterns of Web pages,it can find users' access models automatically and quickly from the vast Web log data,such as frequent access paths,frequent access page groups and user clustering.This will provide foundation for decision making of organizations. Data preprocessing is a key technology in this mining activity.This paper analyses the preprocessing of Web usage mining in detail.After data preprocessing,the number of invalid data can be significantly reduced .**

Keywords:; Web usage mining; data preprocessing; Web log; data mining

I. INTRODUCTION

Web usage mining,also known as Web Log Mining,is the process of extracting interesting patterns in Web access logs[1].Web servers record and accumulate data about user interactions whenever requests for resources are received.Analyzing the web access logs of different web sites can help understand the user behavior and the web structure,thereby improving the website design.Log record has lots of useful information such as URL,IP address,time and so on.Analyzing and discovering log could help us to find more potential users of the web site and trace service quality of the site[2]. Because original log files may exist incomplete or not consistent noisy data,data preprocessmg as necessary before data analysis or pattern mining . It mainly includes data cleaning,user identification,sessions identification,path completion.Effective data preprocessing can improve the quality of mining model and reduce the time needed for mining.

In this paper we discuss different phases of web usage mining and analyze data preprocessmg in detail. Related algorithms and models are given.Based on the experimental evaluation with several web log files and using these algorithms and models we have arrived at a conclusion that the preprocessing is efficient in statistical view.

II. WEB USAGE MINING

The term Web usage mining was introduced by Cooley et al.in 1997 and in accordance with their definition,Web usage mining is the automatic discovery of user access patterns from Web servers[3].A whole process is described in Figure 1.

A. Data Preprocessing

The data preprocessing of Web usage mining is usually complex.Purpose of data preprocessing is to offer structural,reliable and integrated data source to pattern discovery.It consists of four steps: data cleaning, user identification,session identification, path completion[4].

B. Pattern Discovery

In this stage,data mining techniques are used in order to extract patterns of usage from Web data.Pattern discovery is the key process of the Web mining,which covers the algorithms and techniques from several research areas,such as data mining,machine learning,statistics and pattern recognition.The techniques such as statistical analysis,association rules,clustering,classification,sequential pattern and dependency modeling are used to discover rules and patterns[5].OLAP and the data cube structure offer a highly interactive and powerful data retrieval and analysis environment.The knowledge that can be discovered is represented in the form of rules,tables,charts,graphs,and other visual presentation forms for characterizing,comparing,predicting,or classifying data from the Web access log.Visualization can also be used in Web usage mining,and it presents the data in the way that can be understood by users more easily.

C. Pattern Analysis

The final stage of the Web usage mining is pattern analysis.The aim of this process is to extract the interesting rules or patterns from the output of the pattern discovery process by eliminating the irrelative rules or patterns.In addition,OLAP to data cube makes for understanding data from various aspects.Visualization assists an analyst to better apprehend navigation pattern and to predicate trends of data[6].

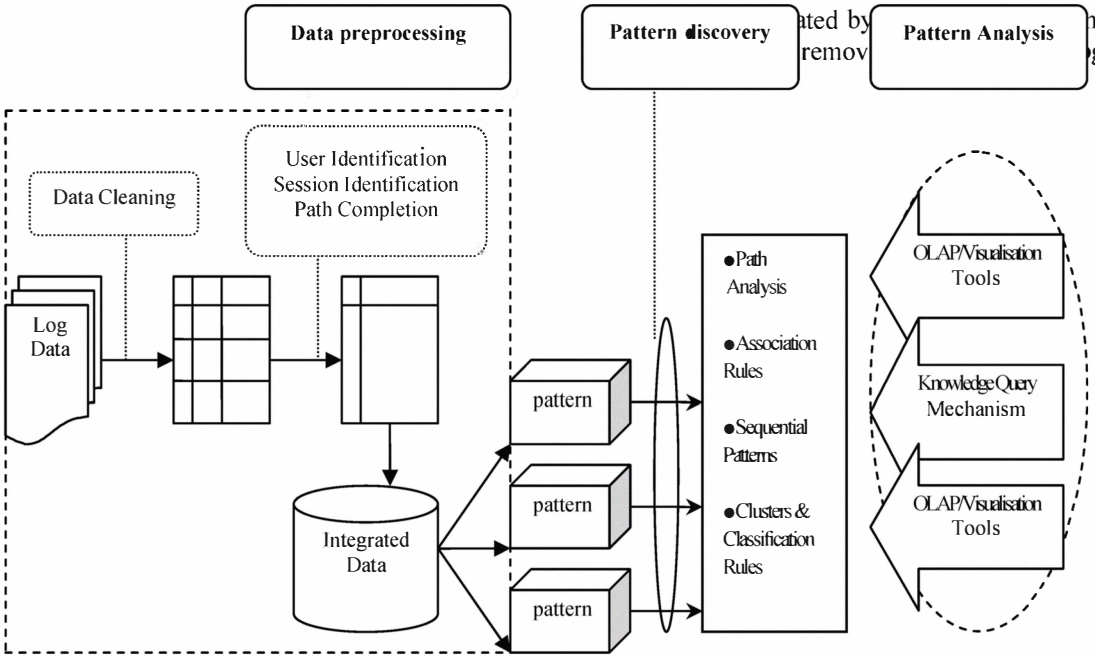

Figure 1. Architecture for Web Usage Mining

III. DATA PREPROCESSING

It is important to understand that the quality data is a key issue when we are going to mining from it.Nearly 80% of mining efforts often spend to improve the quality of data[7]. Web Log files are the best source to know user behavior.But the raw log files contains unnecessary details like image access,failed entries etc.,which will affect the accuracy of pattern discovery and analysis.So preprocessing stage is an important work in mining to make efficient pattern analysis.

A. Data collection

Figure 2. The main sources of Web log files in Web usage mining are (1)Web Servers (2)Web proxy Servers (3)Client browsers.Here data structure of Web logs are shown in Figure 2.

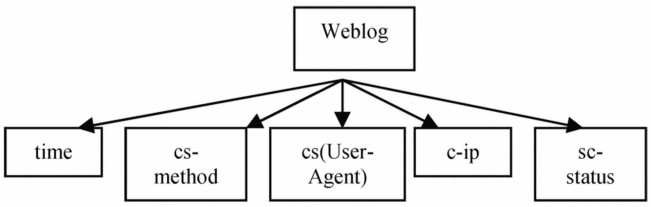

Figure2. Data structure of Web log

B. Data cleaning

Data cleaning is the first step performed in the preprocessing of Web usage mining.The entries which are irrelevant in data analyzing and mining are removed.In data cleaning process,firstly,entries that have status of"error"or"failure"should be removed.Secondly,some access

An algorithm for cleaning the entries of server logs is presented below[8]:

Read record in database.
For each record in database
Read fields(URI–stem)//URI-stem indicates
The target URL//
If fields={*.gif,*.jpg,*.css}then
Remove records
Else
Save records
End if
Next record

C. User identification

User identification means identifying each user accessing website,whose goal is to mine every user's access characteristic,and the make user clustering and provide personal service for the users.But user identification is complicated by the presence of local caches,proxy servers.We assume that each user has unique IP address and each IP address represents one user.But in fact there are three conditions:(1)Some user has unique IP address.(2)Due to proxy server,some user may share one IP address.As of now,we propose following rules for user identification:If there is a new IP address,then there is a new user.If the IP address is same,but the operating system or browser are different,then assumption is that each different agent type for an IP address represents a different user.Moreover we give some notations for user identification.
$Users_i$={User_ID,User_IP,$User\_Ur_l$,User_Time,User_RefferPa

ge,User_Agent},0<i<n where i is the number of total users;User_ID is user's ID have been identified.User_IP is the user's IP address.User_Url is the Web page accessed.User_Time is the time at which user accessed.User_RefferPage is the last page that the user requested.User_Agent is the agent user used.By applying all those above rules,we can easily identify the individual users.

D. Session identification

The goal of session identification is to divide the page requests of each user into individual sessions i.e we find each user's access pattern and frequent path.The best method to identify a session is using a timeout mechanism.

We use the following rules in our experiment to identify individual sessions.(1)If there is a new user,then there is a new session (2)If the referrer page is null in a user session,then we can make sure that there is a new session.(3)If the time between page requests exceeds certain limit(25 or 30 minutes),we can assume that user is starting a new session.Here we propose some notations which help us to identify user's sessions.

$Sessions_i$={User_ID,S_j,[url_{j1},url_{j2},…url_{jk}]},0<i<n where n is the total number of sessions.User_ID stands for user's ID that has been identified;S_j stands for one of the user's sessions;url_{jk} stands for aggregate of Web pages in session S_j.By applying the above stated rules,we can identify user's sessions[9].

E. Path completion

Another critical step in data preprocessing is path completion.There are some reasons that result in path's incompletion,for instance,local cache,agent cache, " post " technique and browser's"back"button can result in some important accesses not recorded in the access log file,and the number of URL recorded in log may be less than the real one.Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log.As a result,the user access paths are incompletely preserved in the Web access log.To discover user's travel pattern,the missing pages in the user access path should be appended.The purpose of the path completion is to accomplish this task.

The better results of data preprocessing,we will improve the mined patterns' quality and save algorithm's running time.It is especially important to Web log files,in respect that the structure of Web log files are not the same as the data in database or data warehouse.They are not structured and complete due to various causations.So it is especially necessary to preprocess Web log files in Web usage mining.Through data preprocessing,Web log can be transformed into another data structure,which is easy to be mined[10].

IV. AN EXAMPLE OF DATA PREPROCESSING APPLICATION

In this paper we took the server log files of site:www.hrbcu.edu.cn.Various analyses have been done to preprocess original Web logs data.

TABLE I. A SINGLE LINE IN THE WEB LOG FILE TAKEN FORMAT

2011-09-23 14:24:18 222.27.186.5 GET /Analytics/Counter.aspx //news.hrbcu.edu.cn 80 -221.212.113.146 Mozilla/4.0(compatible;MSIE 7.0;Windows NT 6.1;Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0; .NET4.0C)

Experimental results of data preprocessing using the forenamed models and algorithms have been in the form of tables.Table2 shows the result of preprocessing statistics on September 23,2011.

TABLE II. DATA PREPROCESSING STATISTICS

Item	Total items in original Web log file	The number of items after data preprocessing	The number of user	The number of session
Amount	12038	4574	825	1027

V. CONCLUSION

Web usage mining is powerful technique used to extract the information from past behavior of users. Data preprocessing plays an important role in this mining activity. The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process, and often requires the use of special algorithms and heuristics not commonly employed in other domains. Web usage mining can be implemented with great efficiency and speed only when invalid data are reduced.

As the complexity of Web applications and user's interaction with these applications increases, the need for intelligent analysis of the Web usage data will also continue to grow. Web usage analysis is used to understand the relationship of user and item which exist in the particular sessions However, without the benefit of deeper domain knowledge, such patterns provide little insight into the underlying reasons for which such items or users are grouped together. Thus, a focus on techniques and architectures for more effective integration and mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications, and more sophisticated tools for Web usage mining that can derive intelligence from user transactions on the Web.

This paper concludes that the proposed algorithms for data preprocessing have been proved efficiency and validity.But it is difficult to take a challenge of over TB level data.It is supposed to adopt different data preprocessing technique according to the characteristic of different Web log files.Related issues need further explanation and research.

REFERENCES

[1] J.Pei,J.Han,B.Mortazavi-asl and Hua Zhu,"Mining access patterns efficiently from web logs,"Knowledge Discovery and Data Mining,2000, vol 1805/2000, pp.396-407

[2] Q.Han, X.Gao,W.Wu,"Study on Web mining algorithm based on usage mining," 9th International Conference on Computer-Aided Industrial Design and Conceptual Design,November,2008

[3] R.Cooley,B.Mobasher,J.Srivastava,"Web mining:information and pattern discovery on the World Wide Web," Tools with Artificial Intelligence,Ninth IEEE International Conference,pp.558–567,USA, November 1997.

[4] K.R Suneetha,R.Krishnamoorthi,"Data preprocessing and easy access retrieval of data through data ware house,"Proceedings of the World Congress on Engineering and Computer Science 2009..

[5] J.Srivasta,R.Cooley, M.Deshpande, P.Tan, "Web usage mining:discovery and applications of usage patterns from Web data,"SIGKDD Explorations.1(2),12-23,2000.

[6] F.Zhang, H.Chang,"Research and Development in Web Usage Mining System-Key Issues and Proposed Solutions:A Survey," Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, November 2002.

[7] David A.Grossman,Ophir Frieder.Information Retrieval:Algorithms and Heuristics(The Information Retrieval Series)(2nd Edition),2004.

[8] Mohd Helmy,Abd Wahab,Nik Shahidah,"Development of Web usage Mining Tools to Analyze the Web Server Logs using Artificial Intelligence Techniques," The 2nd National Intelligence Systems and Information Technology Symposium(ISITS 2007), Malaysia, October 2007.

[9] T.Revathi,M.Praveen Kumar,R.Ravindra Babu,Md.Khaleelur Rahaman,B.Aditya Reddy,"An Effective Analysis of Weblog Files to improve Website Performance," International Journal of Computer Science & Communication Networks,Vol 2(1), pp55-60,2012.

[10] Rajni Pamnani,Pramila Chawan,"Web Usage Mining:A Research Area in Web Mining," Proceedings of ISCET 2010