Volume 4, Issue 8, August 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Usage Mining: Optimize Web Site Navigation Through Log Files

Payal Gupta*
CSE Department of NCCE, ISRANA
India

Dr. Sukhvir Singh
CSE Department of NCCE, ISRANA
India

Abstract— In this paper, we present an overview of web mining and its types used to improve efficiency of web sites as these days for transferring information world wide web has become very popular. We discuss web mining with respect to web data store in web log files. Web mining combines two of the activated research areas: Data Mining and World Wide Web. Web mining is the process of useful information or knowledge from web data. There are three ways to extract knowledge and these are web content, web structure, and web usage data mining. Web usage miming is the area of web miing which deals with the discovery and then analysis of discovered data i.e. usage patterns from web data specifically web logs to improve web based applications. Thus, in this paper we survey the research in web usage mining and will discuss in brief the process of optimization of web sites through web logs.

Keywords— web, data mining, web structure mining, web usage mining, information retrieval, information extraction, web log.

I. INTRODUCTION

The advent of the World Wide Web has caused an increase in the use of the Internet and thus it has become more difficult to access relevant information from the Web The expansion of the World Wide Web has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. World Wide Web is a broadcast medium from where a wide range of information can be obtained very easily and at a low cost . The research focuses on providing effectively and efficiently web service for accessing related web pages because simple structured/query language queries are not adequate to support the increasing demands of today's generation. Data mining is a process of finding hidden information or knowledge in a database .It is also called as data driven discovery, exploratory data analysis.. The web is vast, huge, diverse and dynamic which thus raises the scalability, multimedia data and temporal issues respectively which results in information overload and low precision problem. The growth thus occured in on-line information combined with the unstructured web data necessitates the development of powerful and efficient web data mining tools. Web data mining can be defined as the discovery and analysis of useful information also called as knowledge from the World Wide Web data. Web basically involves three types of data; data on the WWW, the web log data i.e. the data the users who browsed the web pages and the web structure data. Thus, the World Wide Web data mining should focus on three issues; web structure mining, web content mining and web usage mining**.** Web structure mining involves the web document's structures and links mining. Some insight is given on mining structural information on the web. In this paper, we have discussed some applications in web data mining. In this research paper we have traced the client's information which stores in log file and then use this log file to improve the navigation search on web sites. We track each user's ip address, port no, server address etc. after this we removw unwanted data by cleaning the file. A survey of some of the emerging tools and techniques for web usage mining is given in this review paper.

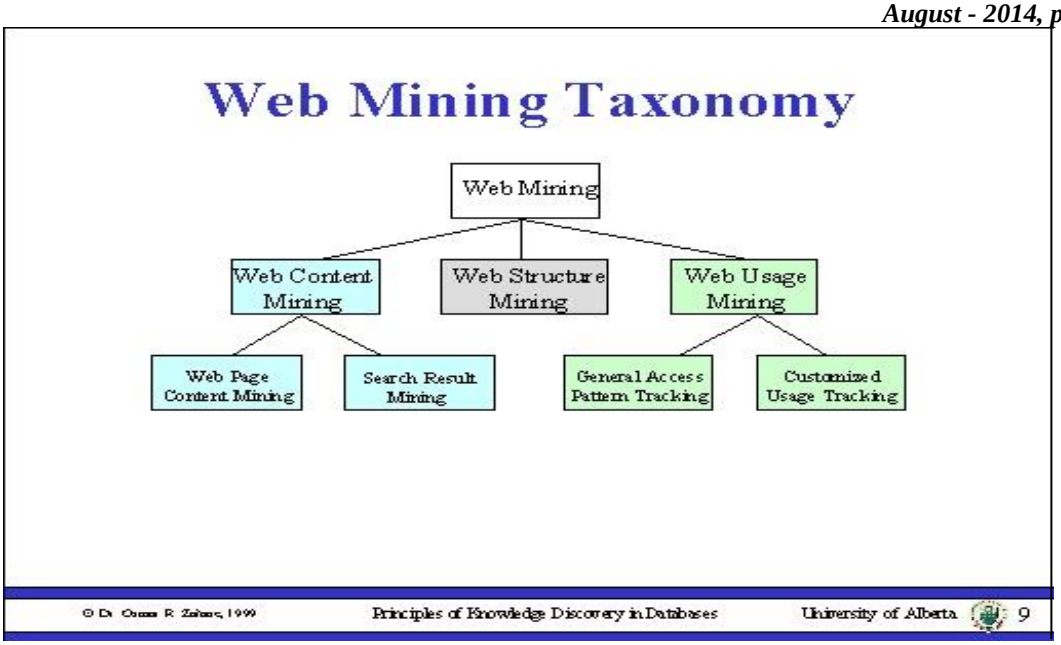
II. WEB MINING

A. Overview

Web mining is defined as a process of using data mining techniques to automatically discover and extract information from the web documents. It basically discovers knowledge from the web data. Today, this area of research is so huge due to the interest of various research communities, the tremendous growth of all the types of information sources are available on the Web. In this, the data is retrieved either online or offline from the different text sources available on the web such as electronic newsletters, newsgroup. We also access the kind of data that is not originally accessible from the web but is accessible such as online text. The information selection and pre-processing step is a process of removing unwanted and incomplete data from the log. The log file contains the information about each activity of users on the web and then use this information to Optimize Website Navigation through Web Server Logs..

B. Web Mining Categories

We categorize web mining into three categories based on which part of the web to mine: Web Content Mining, Web Structure Mining and Web Usage Mining. In this section, thus we give the overview of each of the web category and their detailed explanations are given in their respective sections.
The web minig taxonomy flowchart is defined below.



WEB MINING			
	WEB CONTENT MINING	WEB STRUCTURE MINING	WEB USAGE MINING
View of Data	Unstructured and Semi-Structured	Links Structure	Interactivity
Main Data	Text and Hypertext Documents	Links Structure	Server Logs Browser Logs
Representation	Relational OEM(Edge Labeled Graph)	Graph	Relational Table, Graph

2.1. Web Structure Mining

In web Structure Mining, we are interested in dealing with the structure within the web documents i.e. inter document structure. Web information retrieval tools ignores the valuable information contained in links and makes use of only text pages. Web structure mining helps in generating structural information having summary about web sites and web pages.

2.2. Web content mining

Web content mining involves mining of the web data contents. Web content mining describes the automatic search of information resource available online The Web documents contains several types of data which includes text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured data such as HTML documents, or a more structured data like the data in the tables or database generated HTML pages, The Web content mining is differentiated from two points of view : Information Retrieval View and Database View.

2.3. WEB USAGE MINING

Web usage mining basically focuses on the techniques that predicts the user behavior when the user interacts with the World Wide Web. So whenever new web site is created the main focus is kept on user's interest. It is an activity that involves the automatic discovery of patterns from one or more web servers. The usage data tracks records of each user's activity on web server when the user browses or makes transactions. It is also defined as the discovery of user access patterns from web server logs, which maintains an account of each user. The discovered patterns are represented as collection of objects, pages etc which are accessed by many users having common interest. In this web usage mining, the following activities are observes that are : browsing activities prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site corresponding to the interests of its users. The web usage mining is done in 2 steps:

1. The first step maps the usage data of the world wide web server into the relational data before the data mining technique is performed.

2. The second step uses the data directly by utilizing pre-processing techniques to mine the data.

In general, the data mining methods could be used to mine the usage data after the data have been pre-processed to the desired form.Table Style

III. WEB LOG FILE

Web log files are those files that contains complete information about the user's activity on web server. These log files are created automatically by the corresponding web servers. These log files have text format, most of the times and the size varies from 1KB to 100 MB.

3.1 Web File's Location

Web files are stored at different different locations which are as follows:

I. Web Server Log: These log files are stored by web servers and they donot contain information about visited cached pages.
II. Web Proxy server: Proxy server is used between client and server so that information can be accessed by client easy.
III. Client Browser: In this the web log file is maintained at client sidefor future access.

3.2 Web Log file Types

We have types of log file which are as follows

I. Access Log File: Keeps information about client's request.
II. Error Log File: It keeps record of all the types of error encountered during accessing a specific web page
III. Agent Log File: This keeps complete information about the browser used.

We have different types of file formats of log file **W3C format**: this format is used in the thesis as it is a default log format and can be customized according to the administrators whilw NCSA and IIS log format have fixed size format and cannot be customized.

An algorithm for cleaning the entries of server logs
is presented below –
Input: log_table
Output: refine_log_table
'*' = access pages consist of embedded objects
(i.e .jpg, .gif, etc)
'**' =successful status codes and requested
method(i.e200, GET etc)
Begin
1. Read records in log_table
2. For each record in log_table
3. Read fields (Status code)
4. If Status code='**' , Then Get all fields.
5. If suffix.URL_Link= {*.gif,*.jpg,*.css, *.ico} then
6. Remove suffix.URL_link
7. Save fields in new table.

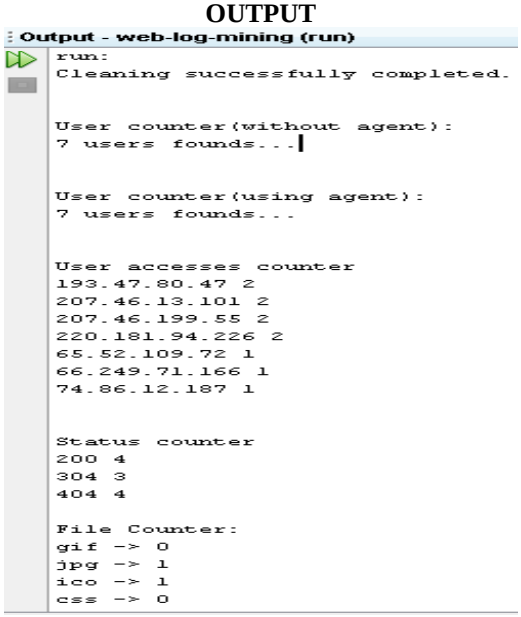
End if
Else
8. Next record
End if
End

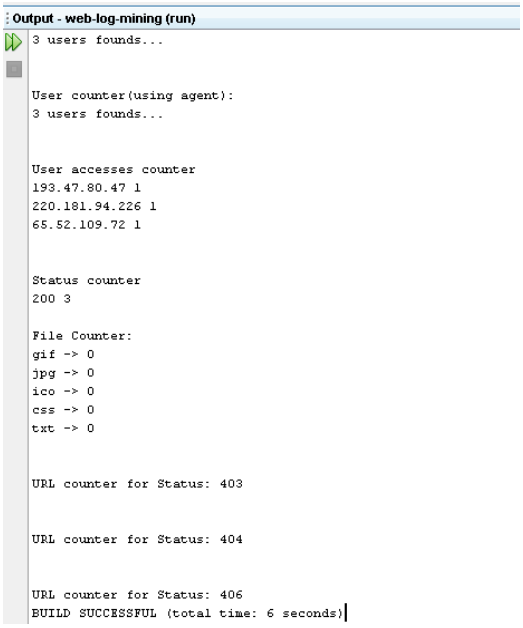
IV. EXPERIMENTAL RESULTS

TABLE 7.1 CONSIDERED FILE

Log File	Size (kb)	Date	No. of Records
A.log	**1154**	**25-05-2014**	**4236**

OUTPUT





Log File Format

Field	Date	Description
Date	date	The date that the activity occurred
Time	time	The time that the activity occurred
Client IP address	c-ip	The IP address of the client that accessed your server
User Name	cs-username	The name of the autheticated user who access your server, anonymous users are represented by -
Servis Name	s-sitename	The Internet service and instance number that was accessed by a client
Server Name	s-computername	The name of the server on which the log entry was generated
Server IP Address	s-ip	The IP address of the server that accessed your server
Server Port	s-port	The port number the client is connected to
Method	cs-method	The action the client was trying to perform
URI Stem	cs-uri-stem	The resource accessed
URI Query	cs-uri-query	The query, if any, the client was trying to perform
Protocol Status	sc-status	The status of the action, in HTTP or FTP terms
Win32 Status	sc-win32-status	The status of the action, in terms used by Microsoft Windows
Bytes Sent	sc-bytes	The number of bytes sent by the server
Bytes Received	cs-bytes	The number of bytes received by the server
Time Taken	time-taken	The duration of time, in milliseconds, that the action consumed
Protocol Version	cs-version	The protocol (HTTP, FTP) version used by the client
Host	cs-host	Display the content of the host header
User Agent	cs(User Agent)	The browser used on the client
Cookie	cs(Cookie)	The content of the cookie sent or received, if any
Referrer	cs(Referrer)	The previous site visited by the user. This site provided a link to the current site

V. CONCLUSION

In this paper, we have discussed some web data mining research issues. . We have defined three types of web data mining in detail. Particularly, we have discussed web data mining with respect to web structure, web content and web usage mining. Data preprocessing is an important task of web usage mining application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. So this system removes accesses to irrelevant items and failed requests in data cleaning. Another task performed in this project is determination of different types of errors that occurred in web surfing. Statistics about hits, page views, visitors are determined. In order to make a website popular among its visitors, System administrator and web designer should try to increase its effectiveness because web pages are one of the most important advertisement tools in international market for business. The obtained results of the study can be used by system administrator or web designer and can arrange their system by determining occurred system errors, corrupted and broken links. In this study, analysis of web server log files has been done. And this result is used to mine optimize website navigation through web server log file.

REFERENCES

[1] Yan Wang ”Web Mining and Knowledge Discovery of Usage Patterns”, 2000.
[2] Srikant and Yang “Mining Web Logs to Improve Website
[3] Kavita Sharma, Gulshan Shrivastava and Vikas Kumar “Web Mining: Today and Tomorrow”, IEEE, 2011
[4] Brijendra Singh and Hemant Kumar Singh, “Web Data Mining Research*:* A Survey”, IEEE, 2010.

