Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Study and Analysis of Web Log Data for Finding the Hidden Facts

Meena Kumari[1], Urooj[2], Prerna[3], J N Shrivastava[4]

[1, 2, 3] M.Tech student of Computer Science and Engineering, Invertis University Bareilly, Uttar Pradesh, India

[4] Associate Professor Department of Computer Science and Engineering, Invertis University Bareilly, Uttar Pradesh, India

Abstract: Web log data is the most important source which contains the information about user id, name, time stamp, IP address, Access request, and number of bytes transferred, URL and user agent. These data are maintained by web servers. Analyzed web log data provide a great idea about the user. The range of web log files is 1 kb to 100kb.There is many interesting patterns are available for web log data. But it is very difficult task to extract the interesting pattern without preprocessed phase. Web log data is generally noisy and ambiguous. Preprocessing is an important process before mining. In this paper we are studying and analyzing web log data for finding the hidden facts using data preprocessing.

Keywords: Data preprocessing, web log file, Data cleaning, web usage mining, session identification, user identification.

I. INTRODUCTION

Data mining is a process of extract data from different perspectives. This process is used to converting scattered data into useful information.

Basically, data mining is the process of finding correlations or patterns among different fields in large relational databases. There are many data mining methods which are used to discover hidden information in the web .Data mining contains the five major Elements.

- Pull out, modify, and load transaction data onto the data warehouse system.
- Keep and manage the data into a multidimensional database system.
- Furnish the data access to business analysts and information technology professionals.
- Appraise the data by application software.
- Show the data in a useful format, such as a graph or table.

Web mining is the application of data mining techniques. The logical structure of Web is a graph structured by documents and hyperlinks, the mining results maybe on Web contents or Web structures. Web mining is used to automatically retrieve and evaluate information for knowledge discovery from web documents .Web mining is categorized into three types.

Web content mining: It deals with the discovery of useful information from the web documents, contents or data. **Web structure mining:** It is process of determining structural information from web. The Structure of web graph consists of web pages as nodes and hyperlinks as edges, connects two web pages /nodes .

Web usage mining: It extract the log data stored in the web server.

Web log file formats are generally designed for debugging purposes, therefore, web accesses are recorded in the order they come.

II. WEB USAGE MINING

Web usage mining referred as web log mining. The web usages mining mainly contains the textual log which are collected by many web server all around the world .There are four stages web usages mining i.e. Data collection, preprocessing ,pattern discovery, pattern analysis .

Data collection

User log files are obtained from various source like server side, client side ,proxy server etc.

Data preprocessing:-Data pre-misprocessing is a necessary steps in the data mining process. It is used to extract useful information from web server log files. .

Pattern discovery phase :-

This is the main phase of web usages mining process. This technique is used for several research area such as data mining, machine learning, pattern recognition applied to the available data. It is also application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on.

Pattern analysis:-
Pattern analysis is the last phase of web usage mining.

Pattern analysis tools
Web site administrators are generally interested in many questions like How are people using the site? Which pages are being accessed most off times? etc. These types of questions are require the analysis of the structure of hyperlinks as well as the contents of the pages. The cease products of such analysis might include:

1. the number of times visits per document,
2. most recent see per document,
3. who has visiting which documents,
4. Most recent use of each hyperlink.

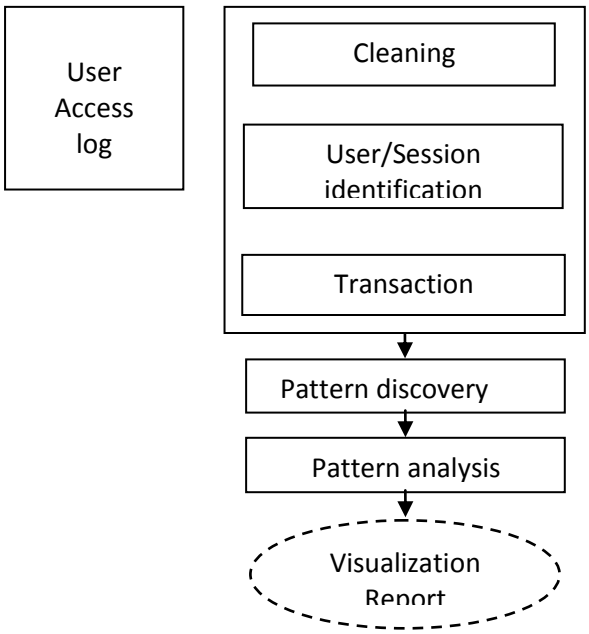

Fig 1: Web usage mining

Web log file format
Web log file is the simple plain text data which contains the information of each user who access the internet.
There are following types of format:

a. W3C Extended log file format.
b. NCSA common log file format
c. IIS log file format.
d. Combined log file format.

W3C (World Wide Web consortium) Extended log file format.
It is default log file format on IIS (Internet Information system) server .In this format the fields are separated by space, time is recorded as GTM (Greenwich Mean Time)
In this format the administrator has the authority to add or remove fields depending on what information want to record. In w3c format of year is YYYY_MM_DD.

```
# Softwacre: Microsoft Internet Information Services7.5

#Version: 1.0

#Date:2013-02-05 06:57:20

#Fields: date time cs-method cs-uri-stem c-ip cs-version sc-status

2013-01-09 3:56:27 GET/WebSite/::1 HTTP/1.1 301
```

Example of W3C log file format

The entry designates on 5 feb,2013,a user with HTTP version 1.1 and GET is the client server method.301 is HTTP status code. Field that are selected but there no information then '-'is placed.

NCSA common log file format:

It stands for National center for Supercomputing Application. It records the basic information about user request, user name, remote host name, date, time, HTTP status code and number of bytes send by server. NCSA is fixed format it cannot customized. Format for year is DD/MM/YYYY. Fields are separated by space, time is local time.

```
::--[19/Jan/2012:10:00:30]"GET/WEBSITE/HTTP/1.1"2001107
```

Obtain information for hidden facts.

Log files contain many parameters that help in recognizing user browsing patterns.
List of parameters is given below

- **User Name:** determines the user who has visited the website and its identification generally is IP address.
- **Visiting Path:** The path taken by the user during website visit is termed as visiting path.
- **Path Traversed:** Path traversed parameter is defined as a path which is traversed by the user within the website.
- **Time Stamp:** The time spent by user on each page is termed as Time stamp. It is also known as session.
- **Page Last Visited:** It deals with page last visited by the user while terminating the website.
- **Success Rate:** Success Rate is computed by downloads and copying activity carried out on the website.
- **User Agent:** User agent is defined as the browser that user uses to send the request to the server.
- **URL:** URL is stands for Uniform Resource Locator which is used as a resource that is accessed by the user .URL may be of any format such as Xml, HTML, and CGI etc.
- **Request Type: Request type is a** method which is used to send the request to the server by the user and it can be either GET or POST method.

User identification

User identification intends determining each user who accesses web site. The purpose of user identification is to obtained each user's access properties, and then make user grouping and provide personal service for each user. We can presume that each user has unique IP address and each IP address represents one user. There are three conditions: for users
(l) Some user has unique IP address.
(2) Some user has two or more IP addresses.
(3), some user may share one IP address. , Due to the existent principle['], we have proposed some rules for user identification:

```
Rule1
IF (IP ADDRESS=new)
Then
{
USER is New
}
Rule 2
IF (IP address is same but Operating System is different)
Then
{
Different user represents same IP address
}
```

For condition (l), these rules lead to accurate result. For condition (2), one user may be assumed as two or more users. But these users have same access pattern, they will be recognized into one group in the mining procedure. For the last condition, those rules can identify some users into one group user, who share one IP address browsing internet. It plays a crucial role for Web usage mining to obtain access pattern of the group user. The group user may be classified into one user respectively. The Web master provides personal service to group user as well .personal user.

III. CONCLUSION

Web sites play an important role for advertisements in international area for universities and other foundation. The quality of a website can be gained by analyzing user accesses of the website by web usage mining. The outcomes of mining can be used to enhance the website design and increase satisfaction which helps in various applications. Log files are the important source to know user behavior. But the raw log files consist of unnecessary details like access of image, failed entries, redirect entries etc., which will affect the accuracy of pattern discovery and analysis. Preprocessing of data is an important stage in mining to make effective and efficient pattern analysis. Therefore, the preprocessing before the data mining in Web logs should become a more important research.

REFERENCE

[1] Theint Theint Aye University of Computer Studies, Mandalay "Web Log Cleaning for Mining of Web Usage Patterns" 978-1-61284-840-2/11/$26.00 ©2011 IEEE.

[2] T. Revathi (Asst. Prof), M. Mohana Rao, Ch. S. Sasanka ,IST & KLCE, K. Jayanth Kumar, B. Uday Kiran IST & KLCE "An Enhanced Pre-Processing Research Framework for Web Log Data"IJARCSSE Volume 2, Issue 3, March 2012 .

[3] FANG YUAN"', LI-JUAN WANG', GE YU' STUDY ON DATA PREPROCESSING ALGORITHM IN WEB LOG MINING International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003

[4] JIANG Chang-bin, Chen LiWeb Log Data Preprocessing Based on Collaborative Filtering, 2010 IEEE DOI 10.1109/ETCS.2010.588118

[5] Nanhay Singh1,et.al COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013 DOI :

