A Study on Web Usage Mining

Dr.V.Kathiresan[1] D.Sridhar[2]
[1]HOD [2]Research Scholar
[1]Department of Computer Applications Engineering
[1,2]Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India

Abstract— In today world Internet usage is growing fast in nowadays and used by all age people with variety of devices. Everyday tasks are handled through internet and user spent more time in the internet to research, entertainment and many other areas. 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. It focuses on the techniques that discover the usage pattern from web log. It uses the data mining techniques like association rule analysis, clustering, classification and machine learning and create new pattern for better user recommendation system for users.

Key words: Web usage mining, Web log mining, Recommender System, Log File formats

I. INTRODUCTION

Web Usage Mining (WUM) is an active field of research and is most likely to generate new knowledge in Internet-based business. WUM applications are being used in some famous websites to understand customer's profiles and their performance in terms of strengths and weaknesses of their website.

A. Data Sources

There are many data sources some are below:

1) Web server logs

Log file registered information about user request history (i.e. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically registered). These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. Server logs do not typically collect user specific information. These files are not usually accessible to general Internet users, only to the webmaster or other administrative entity.

2) Proxy server logs

A web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of web pages as well as the network traffic load on both sides (i.e. server and client). Proxy server logs contain HTTP requests from multiple clients to multiple web servers. This may serve as a data source to discover the usage pattern of an anonymous user groups, sharing a common proxy server.

3) Browser logs

The JavaScript and Java applets used to collect client-side data. This implementation of client-side data collection requires user cooperation, either in enabling the functionality of the JavaScript or Java applets in modified browser.[1]

II. WEB USAGE MINING OVERVIEW

Web usage mining is a web mining technique which is used for discovery and analysis of web usage patterns from web usage data (or web logs). The main aim of web usage mining is to extract interesting information from web logs and, therefore, helps the website administrators to make personalized and adaptive websites that will better serve the needs of the users visiting their websites.[2] The web usage mining process involves three main steps:

1) Preprocessing
2) Pattern Discovery
3) Pattern Analysis

A. Preprocessing

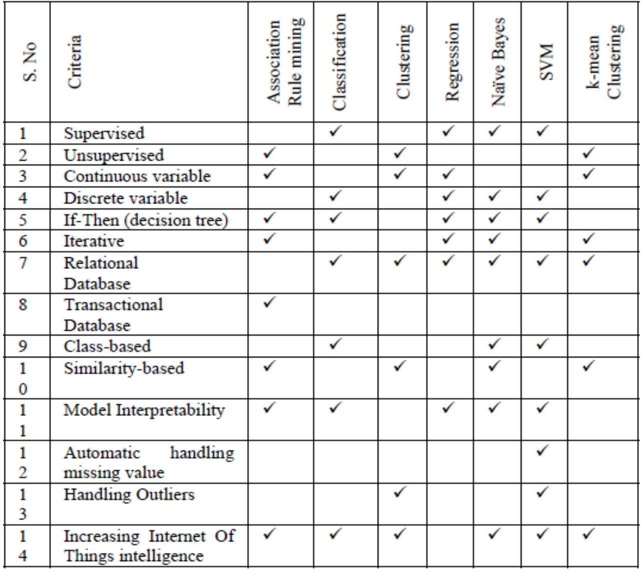
The preprocessing is the first step in web usage mining process in which firstly the web usage log file is cleaned and transformed so as to remove the useless or noisy data from it and to reduce its size. Then using this cleaned log file, user identification (identifying different users through IP address) Preprocessing and session identification (identifying different sessions) is done.[2]

B. Pattern Discovery

Pattern discovery is the second step in web usage mining process in which the cleaned log file generated in the preprocessing step is used to discover web usage patterns.[2]

C. Pattern Analysis

Pattern analysis is the final step in web usage mining process in which the patterns discovered in the second step are further analyzed to generate more interesting patterns and to find more useful information related to the users browsing patterns. [2]



S. No	Criteria	Association Rule mining	Classification	Clustering	Regression	Naïve Bayes	SVM	k-mean Clustering
1	Supervised		✓		✓	✓	✓	
2	Unsupervised	✓		✓				✓
3	Continuous variable	✓		✓	✓			✓
4	Discrete variable		✓		✓	✓	✓	
5	If-Then (decision tree)	✓	✓		✓	✓	✓	
6	Iterative	✓			✓	✓		✓
7	Relational Database		✓	✓	✓	✓	✓	✓
8	Transactional Database	✓						
9	Class-based		✓			✓	✓	
10	Similarity-based	✓		✓		✓		✓
11	Model Interpretability	✓	✓		✓	✓	✓	
12	Automatic handling missing value						✓	
13	Handling Outliers			✓			✓	
14	Increasing Internet Of Things intelligence	✓	✓	✓		✓	✓	✓

Table 1: Comparative Analysis For Mining Algorithms[4]

III. RECOMMENDATION SYSTEM IN ECOMMERCE

Recommender system or Recommendation system is a subclass of information filtering system that intend to foretell the ranking or preference that a web user would assign to an item [3]. The goal of a recommendation system is to help users find personalized information on the web. It is used widely for recommending web pages, movies, articles, restaurants, places to visit, items to buy etc. It learns from user's web navigation behavior and suggests a product or service in which users may be interested. The recommendation system is already used by many major e-commerce web applications to give appropriate suggestion to their consumers. The recommendation could be based on various parameters, such as item popularity, user characteristics such as geographical location or other demographic information such as age, gender and education of users or previous web navigation behavior of most users. The data is the main element of a recommendation system. The data can be obtained by a various ways such as user ratings on products, feedback/reviews from users, etc. The derived data will provide as the base for recommendations to users.

IV. WEB LOG CONTENTS WHICH NEED TO BE ANALYSED

A Web server log file store the information about the requests made to the Web server in a chronological order. The stored web log file contains huge volume of data. Examine each and every type of data is not useful. Thus examining the most pertinent, relevant and useful information in the weblog data may provide more specific information about the patterns for visitors of the web site. This section lists some items that are commonly presented in web log reports. It explains each item, the meaning of the item presented in the report and also the benefits that it potentially brings to the web site from understanding and analyzing it. The raw log files consists of 19 attributes such as Date, Time, Client IP, Auth User, Server Name, Server IP, Server Port, Request Method, URI-Stem, URI-Query, Protocol Status, Time Taken, Bytes Sent, Bytes Received, Protocol Version, Host, User Agent, Cookies, Referrer. These attributes are part of the following logs:

A. Access/Transfer log

The transfer/access log store the detailed information of each request made from user's web browsers to the server. For example, through the host, it can be determined the geographical location of the host, this includes country, state and city in which the viewer of the page is located. This may help in redefining the page for certain population coming from a specific geographical area.

B. Error logs

The error log contains information about errors and failed requests. If the page contains a links to a file does not exist or if the user is unauthorized to access a particular page or file than request may fail.

C. Agent (Browser) log

This log contains the information about the browsers and operating system used by different users to connect the server.

Sometimes, a visitor can access a site from different browsers, counting these visitors differently may help understand the number of visits more precisely

D. Referrer log

When a user access web page by clicking on the link from other site then URL of that site is also recorded in log file which is known as refers log. Or the referrer log contains the URLs of pages on other sites that link to your pages. Further analysis of above variables can generate the following information.

1) Find the user's percentage that accesses the site from a particular domain type (e.g., .com, .edu, .net, .org, .gov). By hits versus accesses analysis.
2) Find the primary clients on the basis of number of hits the server is getting from different user groups.
3) Find the breadth of penetration of the servers. Through number of unique IP addresses visiting the site.
4) Find the optimal time/day by analyzing the quantity of accesses/hits the server receives during specific hours and days of the week. For server maintenance and up gradation.
5) To perform server maintenance and/ or upgrades. Find the average length of a user's session, average time spent by a user on a particular web page, average download times, and navigation pattern of a user through the path analysis of a user on a web site.

The data from Access Logs provides a broad view of a Web servers and users (as indicate by IP address). this type of analysis empowers web server administrators and decision makers to identify potential users of their services and efficient management of their web infrastructure.[3]

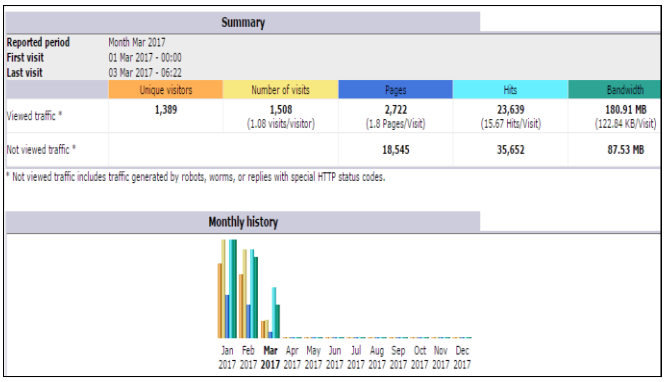
V. LOG FILE FORMATS

Web Site Analyzer can use the information in HTTP, FTP, and other server log files to analyze a site.

The log file formats that Web Site Analyzer can analyze include the following:

- NCSA (Common or Access, Combined, and Separate or 3-Log)
- W3C Extended (used by Microsoft IIS 4.0 and 5.0)
- SunTM ONE Web Server (iPlanet)
- IBM Tivoli Access Manager WebSEAL
- WebSphere Application Server Logs
- FTP Logs
- Custom Log File Format(field information defined by user)

Web log file can be analyzed and different types of reports will be generated like General Statistics, Activity Statistics, Access Statistics, Visitors, Referrers, Browsers, OS, Errors, Tracked Files [6]



Summary

Reported period Month Mar 2017
First visit 01 Mar 2017 - 00:00
Last visit 03 Mar 2017 - 06:22

	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	**1,389**	**1,508** (1.08 visits/visitor)	**2,722** (1.8 Pages/Visit)	**23,639** (15.67 Hits/Visit)	**180.91 MB** (122.84 KB/Visit)
Not viewed traffic *			**18,545**	**35,652**	**87.53 MB**

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

Monthly history

Fig. 1:

VI. Conclusion

Web mining has been used to refer to techniques that help us to find content of web and retrieve the user's interest and needs. Web Usage mining is a complete process to extract Knowledge about browsing behavior of web user from web log. This knowledge is useful in various fields like Website customization, personalization and Recommendation.

References

[1] Mirghani. A. Eltahir, Anour F.A. Dafa-Alla "Extracting Knowledge from Web Server Logs Using Web Usage Mining" International Conference on Computing, Electrical and Electronic Engineering, 2013

[2] Anshul Bhargav, Munish Bhargav, "Pattern Discovery and Users Classification Through Web Usage Mining " International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) IEEE 978-1-4799-4190-2/14, 2014.

[3] Dilip Singh Sisodia, Shrish Verma "Web Usage Pattern Analysis Through Web Logs: A Review" Ninth International Conference on Computer Science and Software Engineering (JCSSE), 2012

[4] Sunena , Kamaljit kaur "Web Usage Mining-Current Trends and Future challenges", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016.

[5] Satyaveer Singh,Mahendra Singh Aswal "Towards a Framework for Web Page Recommendation System based on Semantic Web Usage Mining: A Case Study" 2nd International Conference on Next Generation Computing Technologies (NGCT-2016),2016

[6] https://www.nltechno.com/awstats/awstats.pl?config=destailleur.fr.