Review of Web Usage Mining

Janki Rani
Swami Shraddhanand College, Department of Computer Science,
University of Delhi, Delhi, India

Abstract : Internet has triggered the process of modernization in past few years.. Every aspect of life is highly influenced by internet and its use. Internet has reduced manual labour and enabled to perform many tasks at a mere 'click' of a button. E-commerce, online shopping, Net banking etc are well known applications of Internet. Thus there is need to learn the interaction behavior of web users to better understand their need and accordingly improve the websites. Web usage mining is the area focused on the discovery of navigation pattern of web users. This paper gives an overview of challenges in this emerging area and applications in different fields. Afterwards we have discussed the various sources of data for log mining. A brief description of all the mining techniques has been given through this paper. Finally some software tools along with their characteristics are illustrated which seems to be extremely contributing in expenditure of market and business globally.

Keywords: usage, pattern , mining, web, navigation, server logs

I. INTRODUCTION

Web mining refer to discover useful information from web data and can be applied for better approach of websites. It includes web content mining, web structure mining and web usage mining. Web content mining deals with extracting knowledge from contents displayed on webpages/websites. Web structure mining describes how the contents are organized on a particular webpage/website with the help of hyperlinks. Web structure mining is the process of extracting of usage information from web access logs. Thus also known as Web log mining.

II. WEB USAGE MINING

Web usage mining[1] examines the data generated by user's profiles, user's session and queries imposed by web users. It concentrates on how and when a particular webpage/website has been accessed to increase its usability. Thus it decides the future of e-commerce and hence an important part of web mining.

III. ISSUES OF WEB USAGE MINING

Web is a gold mine of data and it is expanding every fraction of second. Thus it demands tremendous effort to analyze such a huge database, gather usage information, extract irrelevant usage data and then discovering navigation behavior / pattern. The whole process is time consuming and efficient tools are required to carry out the mining task. Privacy and security of user's personal information is a challenging issue.[4]

IV. APPLICATIONS OF USAGE MINING

- It can predict future activities of web users.
- It helps companies to identify their strength to increase their profit.
- It helps in identification of user groups with common interest.
- It provides solutions for improvement of web sites designs.
- It is capable of measuring future marketing efforts by analyzing current pattern.
- It detects web related frauds and thus provides better management.

V. DATA COLLECTION

- ➢ Web Server log : A web server log is a file automatically created by a server to record the accurate browsing behavior of website visitors. These log records the history of visited web pages and tell who accessed a webpage and when. These files are stored in common log format or extended log format. Common log format contains IP address of the site which made the request, login name of the user, full name of the user, date and time of made request, http command, path and filename of the requested file. It is supported by most analysis tools. Extended format gives detailed information than common log format.

- ➢ Web Proxy log : Proxy server is an intermediately server between client and web server. A client connects to a proxy server requesting a file or a service, proxy server then fulfills the request with a reduced level of complexity. Different types of proxy servers perform different tasks. For example, a caching proxy server speed up the loading time by keeping copies of frequently requested resources. It reduces the cost and increases performance.

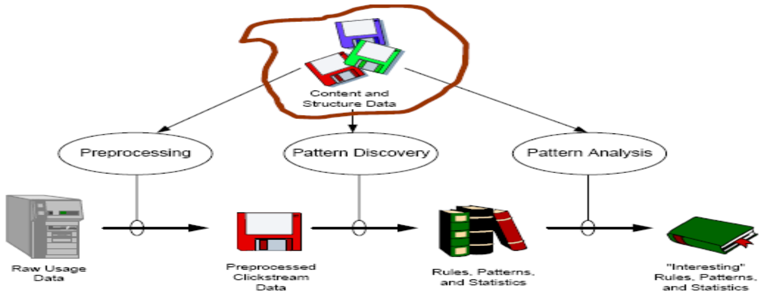
- ➢ Client side logs : Client log participate by using remote agent like java script/java applets or it modifies the source code of the current browser. A modified browser gathers the data about single user over multiple websites[3-4].

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Sample web server log

VI. USAGE MINING TECHNIQUES

As the figure illustrate, there are three major tasks involved in usage mining - Preprocessing, Pattern discovery & Pattern analysis[2-3].


Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Web usage mining process

- Preprocessing[6] : Real world data is generally incomplete & inconsistent which can't produce quality information. Many types of discrepancies exist in data records. Thus elimination of duplicate data entries, filling of incomplete values, correcting user side mistakes is necessary at the earliest level. Subsequently user sessions are identified from user access logs. A session is a series of activities performed by a user during a particular time period. But session detection is not an easy job because multiple factors such as cookies, proxy servers, browsers collectively identify a single session. IP address alone is not sufficient for a user tracing. Finally path detection is done thoroughly as the dynamic structure of web makes it a critical task. User can access the services from local cache or proxy servers also. As a result the actual traversal path remains incomplete in the logs. To better mine the usage of a website these paths must be completed before further processing. Thus the whole method organize the data in such manner so that it can be mined effectively and easily.

- Pattern discovery[5] : Grouping & filtering the user data depending on some characteristics helps to draw some conclusions from integrated log files. We can check last visited pages, frequent visitors & perform other statistical analysis using different variables. Another technique called association rules find relation between pages either accessed by a single user or fetched during a single session. It guides the server to better understand the user requirements and helps pre-fetching or caching. Clustering of pages also contribute in this process by grouping pages with similar/related contents whereas user clustering finds user groups with common interest or similar navigation pattern. Marketing strategies are extremely influenced by visitor’s interest as well as how they surf the sites. Classification of data to a particular category depends upon the specific properties. For example, a bank needs to know which loan customer exists with low, medium or high credit risk or a company wants to categories a product as least, average or most popular in market. Sequential pattern technique predict future navigation pattern by investigating current inter session/transaction patterns. It builds up the foundation for online advertisement and promotion policies.

- Pattern analysis : It requires exhaustive analysis of web structure and contents of a website to find out usage of hyperlinks by web users and their surfing methodology. Knowledge query mechanism such as SQL implements is common method for exploring such facts. Then OLAP (online analytical processing) is another technique used to scrutinize various dimensions of data stored. Market trends are recorded using OLAP tools. Visualization of data is done through depicting graph structures and interpreting different data values according to analysis criteria.

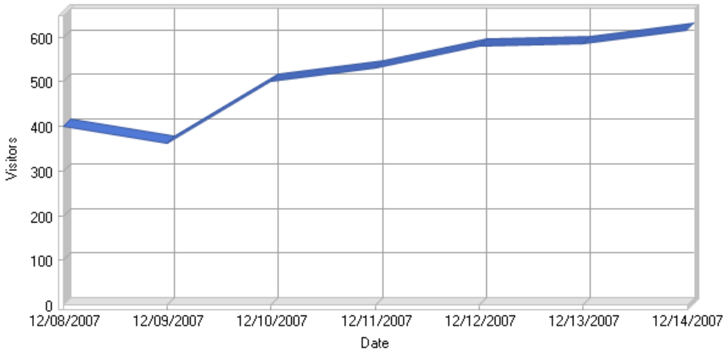
VII. USAGE MINING TOOLS

- Weblog Expert[9] : It is a feature rich access log analyzer that gives information about activity statistics, navigation paths, fetched files, search engines, browsers, operating system etc. It generates reports in both table and chart form. It can convert reports in HTML or PDF formats also. It can even read ZIP compressed log files thus reduces the effort of manual unpacking. Its built in wizards help to quickly create reports for usability analysis of a website. Various filters can be set to explicitly analyze particular web pages or sections of a website.

It can create different types of reports:

- General statistics
- Activity statistics : daily, by hours of the day, by days of week, by weeks and by months
- Access statistics : statistics of pages, files, images, directories, queries etc
- Information about visitors : top-level domain, most active countries/states/cites etc
- Referrers : referring sites, referring URLs, search engines with search keywords
- Error information : error types, detailed error information

Sample reports :

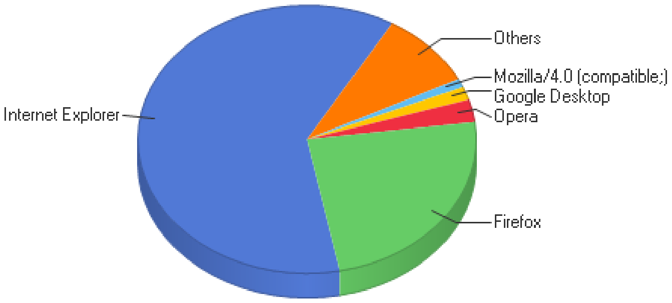


Daily visitors

Hits	
Total Hits	30,474
Visitor Hits	29,191
Spider Hits	1,283
Average Hits per Day	4,353
Average Hits per Visitor	8.18
Cached Requests	3,979
Failed Requests	233
Page Views	
Total Page Views	4,435
Average Page Views per Day	633
Average Page Views per Visitor	1.24

Visitors	
Total Visitors	3,570
Average Visitors per Day	510
Total Unique IPs	2,932
Bandwidth	
Total Bandwidth	567.48 MB
Visitor Bandwidth	548.81 MB
Spider Bandwidth	18.67 MB
Average Bandwidth per Day	81.07 MB
Average Bandwidth per Hit	19.07 KB
Average Bandwidth per Visitor	157.42 KB

General statistics

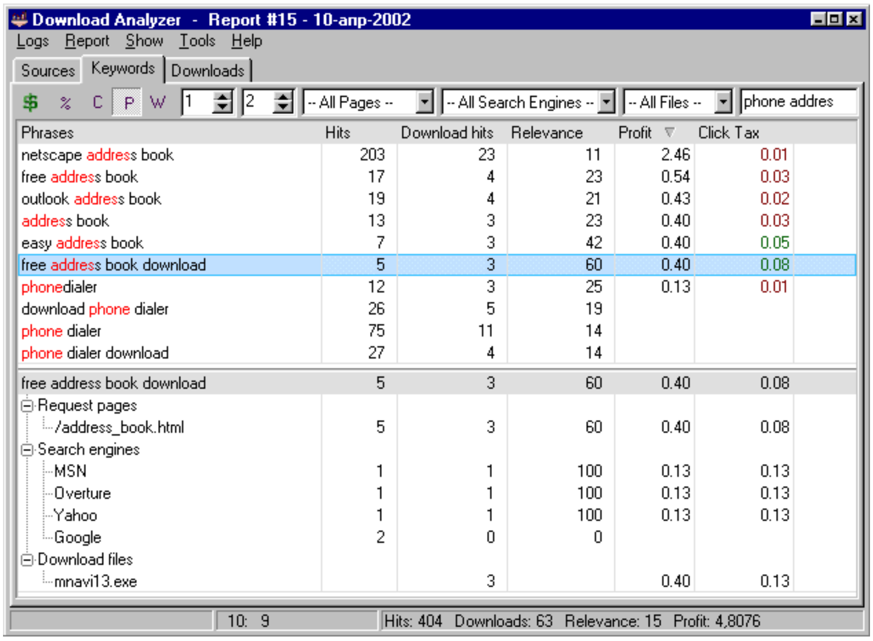


Most used browsers

- ➢ Download analyzer[10] : It is a specialized analyzer for download information of the files. Filters allow relevant qualitative analyses of a site for optimizing the sites for search engines.

The download report contains the following information:

- ▪ Complete downloads - quantity of complete downloads of each file
- ▪ Download hits - quantity of unique visitors to a site who have started downloading file
- ▪ Refer page - the page/site from which the visitor come to your site for download
- ▪ Download page - the page from which the file was directly downloaded



Phrases	Hits	Download hits	Relevance	Profit	Click Tax
netscape address book	203	23	11	2.46	0.01
free address book	17	4	23	0.54	0.03
outlook address book	19	4	21	0.43	0.02
address book	13	3	23	0.40	0.03
easy address book	7	3	42	0.40	0.05
free address book download	5	3	60	0.40	0.08
phonedialer	12	3	25	0.13	0.01
download phone dialer	26	5	19		
phone dialer	75	11	14		
phone dialer download	27	4	14		

free address book download	5	3	60	0.40	0.08
⊟ Request pages					
/address_book.html	5	3	60	0.40	0.08
⊟ Search engines					
MSN	1	1	100	0.13	0.13
Overture	1	1	100	0.13	0.13
Yahoo	1	1	100	0.13	0.13
Google	2	0	0		
⊟ Download files					
mnavi13.exe		3		0.40	0.13

10: 9 Hits: 404 Downloads: 63 Relevance: 15 Profit: 4,8076

Sample download analyzer report

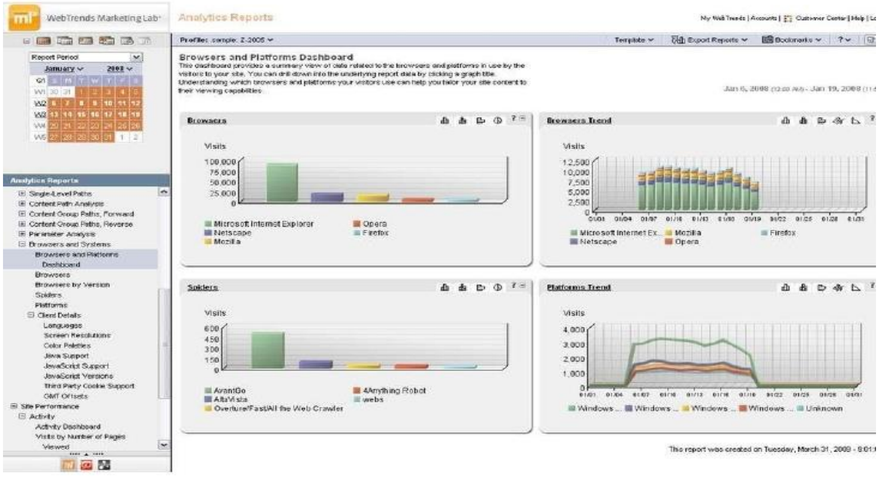
Download analyzer also provides a relevant analysis and can provide :

- Relevance of searched phrases for finding subsequent downloading of files
- Total relevance of group of phrases ,individual pages of a site and search engines
- Average / Total profit for individual pages, group of searched phrases etc

- Web trends[11] : Web trends provides customers and websites information to the companies which helps to decide marketing and advertising strategies. The GUI interface takes the data from web server logs and assemble the information in the form of easy to understand reports. It collects information like new visits, single page visit, duration of visit, referring sites, search engines and searched phrases etc.

Methods of data collection :

- Tracking using scripts : Web trend can implement its own technique by using scripts with each page of website which can record user's information like current visiting webpage, last visited page, IP address, browser, operating system etc
- Web server logs : It can directly fetch the data from server logs.



This dashboard provides a summary view of data related to the browsers and platforms in use by the visitors to your site. You can drill down into the underlying report data by clicking a graph title. Understanding which browsers and platforms your visitors use can help you tailor your site content to their viewing capabilities.

Jan 6, 2008 (12:00 AM) - Jan 19, 2008 (11

This report was created on Tuesday, March 31, 2009 - 9:01:

Sample WebTrend report

- WUM[7] : WUM (Web utilization analysis) tool : It discovers the navigation pattern of visitors to examine the popularity of the website and consequently change the organization of contents or links wherever required. It provides an interesting way to check the parameters in graphical form or tree structure. It provides complete solution for querying and visualization. It first pre-processes data then organizes logs into sessions and finally changes it into a tree structure as per user specified criteria.

- Speed Tracer[8] :It identifies the usage behavior from server logs and create reports either based on user, path or group criteria. Privacy is protected as it does not require user registration details for mining process. It locates user session by tracking navigation paths and generates information about frequent visitors, most accessed web pages, page groups etc. Speed tracer is a very popular web analysis tool to recognize surfing behavior of web user to fix website performance problems.

References :

1. Liu, Bing. *Web data mining*. Springer-Verlag Berlin Heidelberg, 2007.
2. Mobasher, Bamshad. "Web usage mining." *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* 12 (2006).
3. Srivastava, Jaideep, et al. "Web usage mining: Discovery and applications of usage patterns from web data." *ACM SIGKDD Explorations Newsletter* 1.2 (2000): 12-23.
4. Hu, Chen, et al. "World Wide Web usage mining systems and technologies." *Journal of Systemics, Cybernetics and Informatics* 1.4 (2003): 53-59.
5. Omari, Asem. "Web Usage Mining for Adaptive and Personalized Websites." (2011).

6. Pamnani, Rajni, and Pramila Chawan. "Web Usage Mining: A research area in Web mining." *Proceedings of ISCET* (2010): 73-77.
7. Spiliopoulou, Myra, and Lukas C. Faulstich. "WUM: a tool for web utilization analysis." *The World Wide Web and Databases*. Springer Berlin Heidelberg, 1999. 184-203.
8. Wu, K-L., Philip S. Yu, and Allen Ballman. "Speedtracer: A web usage mining and analysis tool." *IBM Systems Journal* 37.1 (1998): 89-105.
9. http://www.weblogexpert.com/
10. http://www.downloadanalyzer.com
11. http://it.toolbox.com/wiki/index.php/WebTrends