National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing (RTMC) 2011
Proceedings published in International Journal of Computer Applications® (IJCA)

Web Usage Mining Systems and Technologies

Sushila Gauthwal
Department of Computer Science
GGSIP University

ABSTRACT
Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications.Web usage mining is used to discover interesting user navigation patterns and can be applied to many real-world problems, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behaviour studies, etc. This article provides a survey and analysis of current Web usage mining systems and technologies. A Web usage mining system performs five major tasks: i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. Each task is explained in detail and its related technologies are introduced. A list of major research systems and projects concerning Web usage mining is also presented, and a summary of Web usage mining is given in the last section.

Keywords World Wide Web, Usage Mining, Navigation Patterns, Usage Data, and Data Mining.

1. INTRODUCTION
World Wide Web Data Mining includes content mining, hyper-link structure mining, and usage mining. All three approaches attempt to extract knowledge from the Web, produce some useful results from the knowledge extracted, and apply the results to certain real-world problems. The first two apply the data mining techniques to Web page contents and hyperlink structures, respectively. The third approach, Web usage mining, the theme of this article, is the application of data mining techniques to the usage logs of large Web data repositories in order to produce results that can be applied to many practical subjects, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behaviour studies, etc.This paper provides a survey and analysis of current Web usage mining technologies and systems. A Web usage mining
system must be able to perform five major functions: i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications.
Requirements of Web Usage Mining
It is necessary to examine what kind of features a Web usage mining system is expected to have in order to conduct effective and efficient Web usage mining, and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to:
Gather useful usage data thoroughly,
Filter out irrelevant usage data,
Establish the actual usage data,
Discover interesting navigation patterns,
Gather useful usage data thoroughly,
Filter out irrelevant usage data,
Establish the actual usage data,
Discover interesting navigation patterns,
Analyze and interpret the navigation patterns correctly, and
Analyze the mining results effectively.

Paper Organization
After many Web usage mining technologies have been proposed and each technology employs a different approach. This article first describes a generalized Web usage mining system, which includes five different functions. Each system function is then explained and analyzed in detail. It is organized as follows:

Section 2 gives a generalized structure of a Web usage mining system and Sections 3 to 7 introduces each of the five system functions and lists its related technologies in turn. Major research systems and projects concerning Web usage mining are listed in Section 8 and the final section summarizes the material Covered in the earlier sections. Related surveys of Web usage mining techniques can also be found in [1 8].

2. SYSTEM STRUCTURE
A variety of implementations and realizations are employed by Web usage mining systems. This section gives a generalized structure of the systems, each of which carries out five major tasks:
Usage data gathering: Web logs, which record user activities on Web sites, provide the most comprehensive, detailed Web usage data.
Usage data preparation: Log data are normally too raw to be used by mining algorithms. This task re-stores the users' activities that are recorded in the Web server logs in a reliable and consistent way.
Navigation pattern discovery: This part of a usage mining system looks for interesting usage patterns contained in the log data. Most algorithms use the method of sequential pattern generation, while the remaining methods tend to be rather ad hoc.
Pattern analysis and visualization: Navigation patterns show the facts of Web usage, but these re-quire further interpretation and analysis before they can be applied to obtain useful results.
Pattern applications: The navigation patterns discovered can be applied to the following major areas, among others: i) improving the page/site design, ii) making additional product or topic recommendations, iii) Web personalization, and iv) learning the user or customer behaviour.

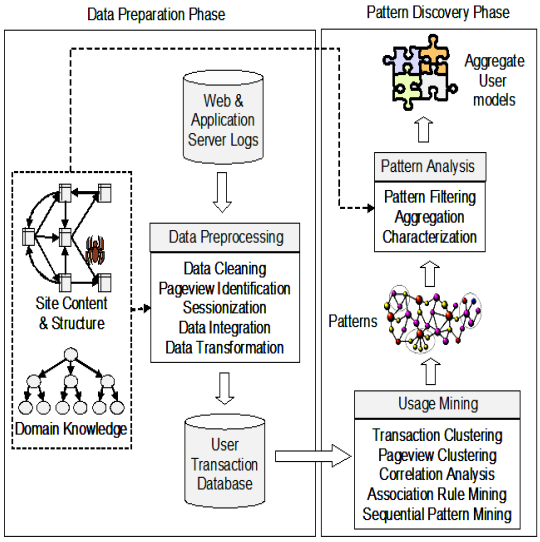

Figure 1: A Web usage mining system structure.

Figure 1 shows a generalized structure of a Web usage mining system; the five components will be detailed in the next five sections. A usage mining system can also be divided into the following two types:

Personal: A user is observed as a physical person, for whom identifying information and personal data/properties are known. Here, a usage mining system optimizes the interaction for this specific individual user, for example, by making product recommendations specifically designed to appeal to this customer.

Impersonal: The user is observed as a unit of unknown identity, although some properties may be accessible from demographic data. In this case, a usage mining system works for a general population, for example, the most popular products are listed for all customers.

This paper concentrates on the impersonal systems. Personal systems are actually a special case of impersonal systems, so readers can easily infer the corresponding personal systems, given the information for impersonal systems.

3. DATA GATHERING

Web usage data are usually supplied by two sources: trial runs by humans and Web logs. The first approach is impractical and rarely used because of the nature of its high time and expense costs and its bias. Most usage mining systems use log data as their data source. This section looks at how and what usage data can be collected.

Web Logs

A Web log file records activity information when a Web user submits a request to a Web server. A log file can be located in three different places: i) Web servers, ii) Web proxy servers, and iii) client browsers, and each suffers from two major drawbacks:

Server-side logs: These logs generally supply the most complete and accurate usage data, but their two draw-backs are:

These logs contain sensitive, personal information, therefore the server owners usually keep them closed.

The logs do not record cached pages visited. The cached pages are summoned from local storage of browsers or proxy servers, not from Web servers.

Proxy-side logs: A proxy server takes the HTTP requests from users and passes them to a Web server; the proxy server then returns to users the results passed to them by the Web server. The two disadvantages are:

Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction.

The request interception is limited, rather than covering most requests.

The proxy logger implementation in Web Quilt [7], a Web logging system, can be used to solve these two problems, but the system performance declines if it is employed because each page request needs to be processed by the proxy simulator.

Client-side logs

Participants remotely test a Web site by downloading special software that records Web us-age or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose. These are pieces of information generated by a Web server and stored in the users' computers, ready for future access. The drawbacks of this approach are:

The design team must deploy the special software and have the end-users install it.

This technique makes it hard to achieve compatibility with a range of operating systems and Web browsers.

Web Log Information

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Examples of the types of information the server preserves in-clude the user's domain, sub domain, and hostname; the re-sources the user requested (for example, a page or an image map); the time of the request; and any errors returned by the server. Each log provides different and various information about the Web server and its usage data. Most logs use the format of a common log file [10] or extended log file... For example, the following is an example of a file recorded in the extended log format.

#Version: 1.0 #Date: 12-Jan-2009 00:00:00 #Fields: time cs-method cs-uri 00:34:23 GET /foo/bar.html

12:21:16 GET /foo/bar.html

12:45:52 GET /foo/bar.html 12:57:34 GET /foo/bar.html

The following list shows the information may be stored in a Web log:

Authuser: Username and password if the server re-quires user authentication.

Bytes: The content-length of the document transferred.

Entering and exiting date Remote IP address or domain name: An IP address is a 32-bit host address defined by the Internet Protocol; a domain name is used to determine a unique Inter-net address for any host on the Internet such as, cs .edu.org. One IP address is usually defined for one domain name, e.g., cs. und. nodak. edu points to 134 .129.216.100.

Modus of request: GET, POST, or HEAD method of CGI (Common Gateway Interface).

Number of hits on the page
Remote log and agent .log
Remote URL.
"request:" The request line exactly as it came from the client.
Requested URL.
rfc931: The remote log name of the user.
Status: The HTTP status code returned to the client, e.g., 200 is "ok" and 404 is "not found."
The CGI environment variables [8] supply values for many of the above items.

4. DATA PREPARATION

The information contained in a raw Web server log does not reliably represent a user session file. The Web usage data preparation phase is used to restore users' activities in the Web server log in a reliable and consistent way. This phase should at a minimum achieve the following four major tasks:
removing undesirable entries
distinguishing among users
building sessions restoring the contents of a session
Removing Undesirable Entries
Web logs contain user activity information, of which some is not closely relevant to usage mining and can be removed with-out noticeably affecting the mining, for example:

All log image entries. The HTTP protocol requires a separate connection for every file that is re-quested from the Web server. Images are automatically downloaded based on the HTML page requested and the downloads are recorded in the logs. In the future, images may provide valuable usage information, but the research on image understanding is still in the early stages. Thus, log image entries do not help the usage mining and can be removed.

Robot assesses
A robot, also known as spider or crawler, is a program that automatically fetches Web pages. Robots are used to feed pages to search engines or other software. Large search engines, like Alta Vista, have many robots working in parallel. As robot-access patterns are usually different from human-access patterns, many of the robot accesses can be detected and removed from the logs. As much irrelevant information as possible should be removed before applying data mining algorithms to the log data.

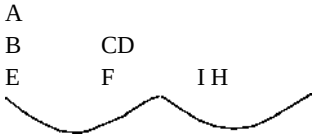

Figure 3: A sample Web site.

Distinguishing among Users

A user is defined as a single individual that accesses files from one or more Web servers through a browser. A Web log sequentially records users' activities according to the time each occurred. In order to study the actual user behaviour, users in the log must be distinguished. Figure 3 is a sample Web site where nodes are pages, edges are hyperlinks, and node A is the entry page of this site. The edges are bi-directional because users can easily use the back button on the browser to return to the previous page. Assume the access data from an IP address recorded on the log are those given in Table 1. Two user paths are identified from the access data: i) A-D- I-H-A-B-F and ii) C-H-B. These two paths are found by heuristics; other possibilities may also exist.

Table 1: Sample access data from an IP address on the site in Figure 3.

No.	Time	Requested URL	Remote URL
1	12:05	A	–
2	12:11	D	A
3	12:22	C	–
4	12:37	I	D
5	12:45	H	C
6	12:58	B	A
7	01:11	H	D
8	02:45	A	–
9	03:16	B	A
10	03:22	F	B

Building Sessions
For logs that span long periods of time, it is very likely that individual users will visit the Web site more than once or their browsing may be interrupted. The goal of session identification is to divide the page accesses of each user into individual sessions. A time threshold is usually used to identify sessions. For example, the previous two paths can be further assigned to three sessions: i) A-D-I-H, ii) A-B-F, and iii) C-H-B if a threshold value of thirty minutes is used.
Restoring the Contents of a Session

This task determines if there are important accesses that are not recorded in the access logs. For example, Web caching or using the back button of a browser will cause information discontinuance in logs. The three user sessions previously identified can be restored to obtain the complete sessions: i) A-D-I-D-H, ii) A-B-F, and iii) C-H-A-B because there are no direct links between I and H and between H and B in Figure 3.

5. NAVIGATION PATTERN DISCOVERY

Many data mining algorithms are dedicated to finding navigation patterns. Among them, most algorithms use the method of sequential pattern generation, while the remaining methods tend to be rather ad hoc.

A Navigation Pattern Example

Before giving the details of various mining algorithms, the following example illustrates one procedure that may be used to find a typical navigation pattern. Assume the following list contains the visitor trails of the Web site in Figure 3.

A-D-I (4)

B-E-F-H (2)

A-B-F-H (3)

A-B-E (2)

B-F-C-H (3)

Where the number inside the parentheses is the number of visitors per trail. An aggregate tree constructed from the list is shown in Figure 4, where the number after the page is the support, the number of visitors having reached the page. A Web

usage mining system then looks for "interesting" navigation patterns from this aggregate tree. Some of the interesting navigation patterns are related to the following three topics: Statistics: for example, which are the most popular paths?

Structure: for example, what pages are usually accessed after users visit page A?

Content: for example, thirty percent of sports page viewers will enter the baseball pages.

Sequential Pattern Generation

The problem of discovering sequential patterns consists of finding intertransaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set ,. The following three systems each use a variant of sequential pattern generation to find navigation pat-terns:

WUM (Web Utilization Miner) [14] discovers navigation patterns using an aggregated materialized view of the Web log. This technique offers a mining language that experts can use to specify the types of patterns they are interested in. Using this language, only patterns having the specified characteristics are saved, while uninteresting patterns are removed early in the process. For example, the following query generates the navigation patterns

select glue(t)

from node B, H

templte B×H as t

where B='B' and H='

MiDAS , extends traditional sequence discovery by adding a wide range of Web-specific features. New domain knowledge types in the form of navigational templates and Web topologies have been incorporated, as well as syntactic constraints and concept hierarchies.

Chen et al. [9] propose a method to convert the original sequence of log data into a set of maximal forward references. Algorithms are then applied to determine the frequent traversal patterns, i.e., large reference sequences, from the maximal forward references obtained.

Ad Hoc Methods Apart from the above techniques of sequential pattern generation, some ad hoc methods worth mentioning are as follows:

Association rule discovery can be used to find unordered correlations between items found in a set of database transactions .. In the context of Web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold.

OLAP (On-Line Analytical Processing) is a category of software tools that can be used to analyze of data stored in a database. It allows users to analyze different dimensions of multidimensional data. For example, it provides time series and trend analysis views. WebLogMiner uses the OLAP method to analyze the Web log data cube, which is constructed from a database containing the log data. Data mining methods such as association or classification are then applied to the data cube to predict, classify, and discover interesting patterns and trends. Büchner and Mulvenna [7] also make use of a generic Web log data hypercube. Various online analytical Web usage data mining techniques are then applied to the hypercube to reveal marketing intelligence.

Borges and Levene's . model views navigation records in terms of a hypertext probabilistic grammar, which is a probabilistic regular grammar. For this grammar, each non-terminal symbol corresponds to a Web page and a production rule corresponds to a link between pages. The higher probability generated strings of the grammar correspond to the user' s preferred trails.

Pei et al. propose a data structure WAP-tree to store highly compressed, critical information contained in Web logs, together with an algorithm WAP-mine that is used to discover access patterns from the WAP-tree.

6. PATTERN ANALYSIS AND VISUALIZATION

Navigation patterns, which show the facts of Web usage, need further analysis and interpretation before application. The analysis is not discussed here because it usually requires human intervention or is distributed to the two other tasks: navigation pattern discovery and pattern applications. Navigation patterns are normally two-dimensional paths that are difficult to perceive if a proper visualization tool is not supported. A useful visualization tool may provide the following functions:

•Displays the discovered navigation patterns clearly.

•Provides essential functions for manipulating navigation patterns, e.g., zooming, rotation, scaling, etc.WebQuilt allows captured usage traces to be aggregated and visualized in a zooming interface. The visualization also shows the most common paths taken through the Web site for a given task, as well as the optimal path for that task as designated by the designers of the site.

7. PATTERN APPLICATIONS

The results of navigation pattern discovery can be applied to the following major areas, among others: i) improving site/page design, ii) making additional topic or product recommendations, iii) Web personalization, and iv) learning user/customer behaviour. Web caching, a less important application for navigation patterns, is also discussed.

Web Site/Page Improvements

The most important application of discovered navigation patterns is to improve the Web sites/pages by (re)organizing them. Other than manually (re)organizing the Web sites/pages ., there are some other automatic ways to achieve this. Adaptive Web sites [26] automatically improve their organization and presentation by learning from visitor access patterns. They mine the data buried in Web server logs to produce easily navigable Web sites. Clustering mining and conceptual clustering mining techniques are applied to synthesize the index pages, which are central to site organization.

Additional Topic or Product Recommendations

Electronic commerce sites use recommender systems or collaborative filtering to suggest products to their customers or to provide consumers with information to help them decide which products to purchase. Various technologies have been proposed for recommender systems. and many electronic commerce sites have employed recommender systems in their sites. For further studies, the Group Lens research group [16] at the University of

Minnesota is known for its successful projects on various recommender systems.

8. WEB PERSONALIZATION

Web personalization (re)organizes Web sites/pages based on the Web experience to fit individual users' needs. It is a broad area that includes adaptive Web sites and recommender systems as special cases. The WebPersonalizer system [23] uses a subset of Web log and session clustering techniques to derive usage profiles, which are then used to generate recommendations. An overview of approaches for incorporating se-mantic knowledge into the Web personalization process is given in the article by Dai and Mobasher [12].

7.2 User Behaviour Studies

Knowing the users' purchasing or browsing behaviour is a critical factor for the success of E-commerce. The 1: 1Pro system . constructs personal profiles based on customers' transactional histories. The system uses data mining techniques to discover a set of rules describing customers' behaviour and supports human experts in validating the rules. Fu et al. . propose an algorithm to cluster Web users based on their access patterns, which are organized into sessions representing episodes of interaction between Web users and the Web server. Using attributed-oriented induction, the sessions are then generalized according to the page hierarchy, which organizes pages according to their generalities. The generalized sessions are finally clustered using a hierarchical clustering method.

7.3 Web Caching

Another application worth mentioning is Web caching, which is the temporary storage of Web objects (such as HTML documents) for later retrieval. There are significant advantages to Web caching, e.g., reduced bandwidth consumption, reduced server load, and reduced latency. Together, they make the Web less expensive and improve its performance . Web caching may in turn be enhanced by navigation patterns. Lan et al. [12] propose an algorithm to make Web servers "pushier." Which document is to be prefetched is determined by a set of association rules mined from a sample of the access log of the Web server. Once a rule of the form "Document1 > Doc ument2" has been identified and selected, the Web server decides to prefetch "Document2" if "Document1" is requested. Two use the method of sequential pattern generation, while the rest tend to use ad hoc methods. Sequential pattern generation does not dominate the algorithms, since navigation patterns are defined differently from one application to another and each definition may require a unique method.

Table 2: Major research systems and projects concerning Web usage mining

	Title	URL	Major
1	Adaptive Web	http://www.cs.washington.edu/research/adaptive/	Pattern
2	Group Lens	http://www.cs.umn.edu/Research/GroupLens/	Recom
3	MMAS		Sequence
4	WebQu	http://guir.berkeley.edu/pro	Prox
5	WebLo	http://www.dbminer.com/	OL
6	WebSif	http://www.cs.umn.edu/Re	Data
7	WUM	http://wum.wiwi.hu-berlin.de/	Sequence discover

9. REFERENCES

[1] Access log analyzers. Retrieved June 02, 2003 from http://www.uu.se/Software/Analyzers/Accessanalyzers.html

[2] Gediminas Adomavicius and Alexander Tuzhilin. Using data mining methods to build customer profiles.

[3]IEEE Computer, 34(2):74-82, February 200] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Proceeding of the 20th Very Large DataBases Conference (VLDB),pages 487-499, Santiago, Chile, 1994.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering, pages 3-14,Taipei, Taiwan, March 1995.

[5] José Borges and Mark Levene. Data mining of user navigation patterns. In Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD),pages 31-36, San Diego, California, August 1999.

[6] Alex G. Büchner, Matthias Baumgarten, Sarabjot S.Anand, Maurice D. Mulvenna, and John G. Hughes.Navigation pattern discovery from Internet data. In Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD), San Diego, California, August 1999.

[7] Alex G. Büchner and Maurice D. Mulvenna. Discovering Internet marketing intelligence through online analytical Web usage mining. ACM SIGMOD Record,27(4):54-61, December 1998.

[8] CGI environment variables. Retrieved May 15, 2003 From http://hoohoo.ncsa.uiuc.edu/cgi/env.html

[9] Ming-Syan Chen, Jong Soo Park, and Philip S. Yu.Efficient data mining for path traversal patterns.

IEEE Transactions on Knowledge and Data Engineering,8(6):866-883, 1996.

[10] Common log file format. Retrieved June 02, 2003 from http://www.w3.org/Daemon/User/Config/Logging.html

SYSTEMICS, CYBERNETICS AND INFORMATICS VOLUME 1 - NUMBER 4

[11] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1(1):5-32, February 1999.

[12] Bin Lan, Stephane Bressan, and Beng Chin Ooi. Making Web servers pushier. In Proceedings of the Workshop on Web Usage Analysis and User Profiling, pages 112-125, San Diego, California, August 1999.

[13] Myra Spiliopoulou and Lukas C. Faulstich. WUM: A tool for Web utilization analysis. In Proceedings of the Workshop on the Web and Databases (WEBDB),pages 184-203, Valencia, Spain, March 1998.

[14] Extended log file format. Retrieved June 03, 2003 from http://www.w3.org/TR/WD-logfile.html

[15] Yongjian Fu, Kanwalpreet Sandhu, and Ming-Yi Shih. A generalization-based approach to clustering of Web usage sessions. In Brij M. Masand and Myra Spiliopoulou, editors, *Web Usage Analysis and User Profiling,Lecture Notes in Artificial Intelligence*, 1836:21-38,Springer, 2000.

[16] GroupLens Research. Retrieved May 12, 2003 from http://www.cs.umn.edu/Research/GroupLens/

[17] Jason I. Hong and James A. Landay. WebQuilt: A framework for capturing and visualizing the Web experience. In *Proceedings of the 10th International World Wide Web Conference*, pages 717-724, Hong Kong, 2001.

[18] Wen-Chen Hu, Xuli Zong, Hung-Ju Chu, and Jui-Fa Chen. Usage mining for the World Wide Web. In *Proceedings of the 6th World Multi-Conference on Systemics, Cybernetics and Informatics* (*SCI*), pages 75-80, Orlando, Florida, July 14-18, 2002.

[19] Melody Y. Ivory and Marti A. Hearst. Improving Web site design. *IEEE Internet Computing*, 6(2):56-63, March/April 2002.

[20] Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2(1):1-15,2000.

[21] [32] Jaideep Srivastava, Robert Cooley, Mukund Deshpande,and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from Web data. *ACM Special Interest Group on Knowledge Discovery and Data Mining* (*SIGKDD*) *Explorations*, 1(2):12-23,2000.