International Journal of Instrumentation Control and Automation

Volume 1 | Issue 1 Article 5

April 2011

Web Usage Mining: A Survey on Pattern Extraction from Web Logs

S. K. Pani
P.G. Department Of Computer Science, RCMA; Bhubaneswar, Orissa, India, Subhendu_pani@rediffmail.com

L. Panigrahy
Department of Computer Science and Engineering; Konark Institute of Science and Technology; Bhubaneswar, Orissa, India, mynamelingaraj@gmail.com

V.H. Sankar
Department of Computer Science and Engineering; Konark Institute of Science and Technology; Bhubaneswar, Orissa, India, himashankar.V@gmail.com

Bikram Keshari Ratha
P.G. Department Of Computer Science, Utkal University,Bhubaneswar, Orissa, India, vkramus@gmail.com

S.K. Padhi
Department of Computer Science and Engineering; Konark Institute of Science and Technology; Bhubaneswar, Orissa, India, Sanjaya2004@yahoo.com

Recommended Citation

Pani, S. K.; Panigrahy, L.; Sankar, V.H.; Ratha, Bikram Keshari; Padhi, S.K.; and Mandal, A.K. (2011) "Web Usage Mining: A Survey on Pattern Extraction from Web Logs," *International Journal of Instrumentation Control and Automation*: Vol. 1 : Iss. 1 , Article 5.
DOI: 10.47893/IJICA.2011.1004
Available at: https://www.interscience.in/ijica/vol1/iss1/5

Web Usage Mining: A Survey on Pattern Extraction from Web Logs

Authors
S. K. Pani, L. Panigrahy, V.H. Sankar, Bikram Keshari Ratha, S.K. Padhi, and A.K. Mandal

Web Usage Mining: A Survey on Pattern Extraction from Web Logs

[1]S. K. Pani, , [2]L. Panigrahy, [2]V.H.Sankar, [3]Bikram Keshari Ratha, [2]A.K.Mandal, [2]S.K.Padhi
[1] P.G. Department Of Computer Science, RCMA; Bhubaneswar, Orissa, India
[2] Department of Computer Science and Engineering; Konark Institute of Science and Technology; Bhubaneswar, Orissa, India
[3]P.G. Department Of Computer Science, Utkal University,Bhubaneswar, Orissa, India
E-mail: Subhendu_pani@rediffmail.com; mynamelingaraj@gmail.com; Himashankar.V@gmail.com; vkramus@gmail.com; Sanjaya2004@yahoo.com

***Abstract*— As the size of web increases along with number of users, it is very much essential for the website owners to better understand their customers so that they can provide better service, and also enhance the quality of the website. To achieve this they depend on the web access log files. The web access log files can be mined to extract interesting pattern so that the user behaviour can be understood. This paper presents an overview of web usage mining and also provides a survey of the pattern extraction algorithms used for web usage mining.**

***Keywords*— web mining, pattern extraction, usage mining, preprocessing**

I. INTRODUCTION

In this world of Information Technology, accessing information is the most frequent task. Every day we have to go through several kind of information that we need and what we do? Just browse the web and the desired information is with us on a single click. Today, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customer 24X7. On the other side visitors are also availing those facilities.

In the last fifteen years, the growth in number of web sites and visitors to those web sites has increased exponentially. The number of users by June 30 2010 was 1,966,514,816[18] which is 28.7% of the world's population. The number of active web sites is *125,522,259* [19] as on 13-Dec-2010. Due to this growth a huge quantity of web data has been generated. To mine the interesting data from this huge pool, data mining techniques can be applied. But the web data is unstructured or semi structured. So we can not apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is used to discover interest patterns which can be applied to many real world problems like improving web sites, better understanding the visitor's behavior, product recommendation etc.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services (Etzioni,1996). Web mining is categorized into 3 types. 1. Content Mining (Examines the content of web pages as well as results of web Searching) 2. Structure Mining (Exploiting Hyperlink Structure) 3. Usage Mining (analyzing user web navigation)

Web usage mining is a process of picking up information from user how to use web sites. Web content mining is a process of picking up information from texts, images and other contents. Web structure mining is a process of picking up information from linkages of web pages.

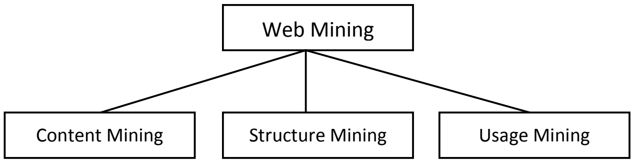

Figure 1: Web mining classification

TABLE 1

THE RELATIONSHIP AMONG THE DIFFERENT AREAS OF WEB MINING

Type	Structure	Form	Object	Collection
Usage	Accessing	Click	Behavior	Logs
Content	Pages	Text	Index	Pages
Structure	Map	Hyperlinks	Map	Hyperlinks

These 3 approaches attempts to extract knowledge from Web generate some useful result from that knowledge and apply the result to certain real world problems. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files.

Web usage mining is one of the prominent research area due to these following reasons. a) One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining. b) Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses. Prefetching and caching policies can be made on the basis of frequently accessed pages to improve latency time. c) Common access behaviors of the users can be used to improve the actual design of web pages and for making other modifications to a Web site. d) Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

Five major steps followed in web usage mining are

1. Data collection – Web log files, which keeps track of visits of all the visitors
2. Data Integration – Integrate multiple log files into a single file
3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern
6. Pattern applications – Apply the pattern in real world problems

II. RELATED WORK ON PATTERN EXTRACTION

The existing research works done by different authors can be categorized into a) Association Rule Mining (ARM), b) Clustering, c) Classification

TABLE 2

THE RESEARCH WORK DONE ON ARM

Algorithm Used	Authors	Year
Maximal forward references	Ming-Syan Chen, Jong Soo Park, Philip S. Yu	1998
Markov Chains	Jianhan Zhu, Jun Hong, and John G. Hughes	2002
Improved AprioriAll	WANG Tong, HE Pi-lian	2005
Fpgrowth and Prefixspan	Hengshan Wang, Cheng Yang, Hua Zeng	2006
Custom Built APRIORI Algorithm	Sandeep Singh Rawat, Lakshmi Rajamani	2010

Ming-Syan et al., [6] proposed a new data mining algorithm that involves mining path traversal patterns in a distributed information-providing environment where documents or objects are linked together to facilitate interactive access. Their solution procedure consists of two steps. First, derive an algorithm to convert the original sequence of log data into a set of maximal forward references. Second, derive another algorithms to determine the frequent traversal patterns i.e., large reference sequences from the maximal forward references obtained. Jianhan Zhu et al., [7] applied the Markov chains to model user navigational behavior. They proposed a method for constructing a Markov model of a web site based on past visitor behavior. Then the Markov model is used to make link predictions that assist new users to navigate the Web site. WANG Tong et al., [8] offers an improved algorithm based on the original AprioriAll algorithm. The new algorithm adds the property of the UserID during the every step of producing the candidate set and every step of scanning the database by which to decide whether an item in the candidate set should be put into the large set which will be used to produce next candidate set. Hengshan Wang et al., [9] introduced two prevalent data mining algorithms - FPgrowth and PrefixSpan into WUM. Maximum Forward Path (MFP) is also used in the web usage mining model during sequential pattern mining along with PrefixSpan so as to reduce the interference of "false visit" caused by browser cache and raise the of mining frequent traversal paths. Sandeep Singh Rawat et al., [13] proposed a custom-built apriori algorithm which is based on the old Apriori algorithm, to find the effective pattern analysis

TABLE 3
THE RESEARCH WORK DONE ON CLUSTERING

Algorithm Used	**Authors**	**Year**
Self Organized Maps	Paola Britos, Damián Martinelli, Hernán Merlino, Ramón García-Martínez	2007
Graph Partitioning	Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman	2008
Ant-based	Kobra Etminani, Mohammad-R. Akbarzadeh-T, Noorali Raeeji Yanehsari	2009
K-mean with Genetic Algorithm	N. Sujatha, K. Iyakutty	2010

Paola Britos et al., [10] described the capacity of use of Self Organized Maps, kind of artificial neural network, in the process of Web Usage Mining to detect user's patterns. The process detail the transformations necessaries to modify the data storage in the Web Servers Log files to an input of Self Organized Maps. Mehrdad Jalali et al., [11] presented an approach which is based on the graph partitioning for modeling user navigation patterns. In order to mining user navigation patterns, they establish an undirected graph based on connectivity between each pair of the web pages and also proposed novel formula for assigning weights to edges of the graph. Kobra Etminani et al., [12] applied ant-based clustering algorithm to pre-processed logs to extract frequent patterns for pattern discovery and then it is displayed in an interpretable format. N. Sujatha et al., [15] have proposed a new framework to improve the web sessions' cluster quality from k-means clustering using Genetic Algorithm (GA). Initially a modified k-means algorithm is used to cluster the user sessions. The refined initial starting condition allows the iterative algorithm to converge to a "better" local minimum. And in the second step, they have proposed a GA based refinement algorithm to improve the cluster quality.

TABLE 4
THE RESEARCH WORK DONE ON CLASSIFICATION

Algorithm Used	Authors	Year
Naive Bayesian	Mahdi Khosravi, Mohammad J. Tarokh	2010

Mahdi Khosravi et al., [14] proposed a novel approach for dynamic mining of users' interest navigation patterns, using naïve Bayesian method

III. BASIC CONCEPTS

A. Data Mining

There are two classes [4] of data mining namely i) to summarize or characterize general properties of data in repository which is called Descriptive and ii) to perform inference on current data, to make predictions based on the historical data which is called Prescriptive. There are various data mining techniques available which also can be applied to web data mining. Few techniques are listed below.

1) Association Rules Mining: When the book Data Mining Concepts and Techniques is bought, 40% of the time the book Database System is bought together, and 25% of the time the book Data Warehouse is bought together. Those rules discovered from the transaction database of the book store can be used to rearrange the way of how to place those related books, which can further make those rules more strong

2) Sequential Pattern Mining: Association rule mining does not take the time stamp into account, the rule can be Buy A=>Buy B. If we take time stamp into account then we can get more accurate and useful rules such as: Buy A implies Buy B within a week, or usually people Buy A every week. As we can see with the second kind of rules, business organizations can make more accurate and useful prediction and consequently make more sound decisions. A database consists of sequences of values or events that change with time, is called a time-series database [Han and Kamber 2000], a time-series database records the valid time of each dataset. For example, in a time-series database that records the sales transaction of a supermarket, each transaction includes an extra attribute indicate when the transaction happened. Time-series database is widely used to store historical data in a diversity of areas such as, financial data, medical data, scientific data and so on. Different mining techniques have been designed for mining time-series data, basically there are four kinds of patterns we can get from various types of time-series data:1) Trend analysis, 2) Similarity search, 3) Sequential patterns and 4) Periodical patterns. Sequential patterns: sequential pattern mining is trying to find the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences. We can find the sequential patterns of specific individual items; also we can find the sequential patterns cross different items. Sequential pattern mining is widely used in analyzing of DNA sequence. An example of sequential patterns is that every time Microsoft stock drops 5%, IBM stock will also drops at least 4% within three days.

3) Classification: Classification is to build (automatically) a model that can classify a class of objects so as to predict the classification or missing attribute value of future objects (whose class may not be known). It is a two-step process. In the first process, based on the collection of training data set, a model is constructed to describe the characteristics of a set of data classes or concepts. Since data classes or concepts are predefined, this step is also known as supervised learning (i.e., which class the training sample belongs to is provided). In the second step, the model is used to predict the classes of future objects or data. A decision tree for the class of buy laptop, indicate whether or not a customer is likely to purchase a laptop. Each internal node represents a decision based on the value of corresponding attribute, also each leaf node represents a class (the value of buy laptop=Yes or No). After this model of buy laptop has been built, we can predict the likelihood of buying laptop based on a new customer's attributes such as age, degree and profession. That information can be used to target customers of certain products or services, especially widely used in insurance and banking.

4) Clustering: Classification can be taken as supervised learning process, clustering is another mining technique similar to classification. However clustering is a unsupervised learning process. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. In classification which record belongs which class is predefined, while in clustering there is no predefined classes. In clustering, objects are grouped together based on their similarities. Similarities between objects are defined by similarity functions, usually similarities are quantitatively specified as distance or other measures by corresponding domain experts. For example, based on the expense, deposit and draw patterns of the customers, a bank can clustering the market into different groups of people. For different groups of market, the bank can provide different kinds of loans for houses or cars with different budget plans. In this case the bank can provide a better service, and also make sure that all the loans can be reclaimed.

B. Log Files

In order to manage a web server effectively, it is necessary to get feedback about the activity and performance of the server as well as any problems that may be occurring. Web server creates and maintains log files for this purpose [3]. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site.

1) Log Formats: W3C maintains a standard format for web server log files, but other proprietary formats exist. For example IIS provides six different log file formats which are used to track and analyze information about IIS-based sites and services such as 1. W3C Extended Log File Format, 2. W3C Centralized Logging, 3. NCSA Common Log File Format, 4. IIS Log File Format, 5. ODBC Logging, 6. Centralized Binary Logging. In addition to the six available formats, custom log file format can also be configured.

A log file in the W3C extended format contains a sequence of lines containing ASCII characters. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by white space. If a field is unused in a particular entry dash "-" marks the omitted field. Directives record information about the logging process itself. Lines beginning with the # character contain directives. The following directives are defined in the W3C Extended format [16, 17]:

The following is an example of a record in the extended log format that was produced by the Microsoft Internet Information Server (IIS):

```
#Software: Microsoft Internet Information
Server 4.0
#Version: 1.0
#Date: 1998-11-19 22:48:39
#Fields: date time c-ip cs-username s-ip
cs-method cs-uri-stem cs-uri-query sc-
status sc-bytes cs-bytes time-taken cs-
version cs(User-Agent) cs(Cookie)
cs(Referrer)

1998-11-19 22:48:39 206.175.82.5 -
208.201.133.173 GET
/global/images/navlineboards.gif - 200
540 324 157 HTTP/1.0
Mozilla/4.0+(compatible;+MSIE+4.01;+Windo
ws+95) USERID=CustomerA;+IMPID=01234
http://yourturn.rollingstone.com/webx?98@
@webx1.html
```

Description of headers

c	Client
s	Server
r	Remote
cs	Client to Server.
sc	Server to Client.
sr	Server to Remote Server, this prefix is used by proxies.
rs	Remote Server to Server, this prefix is used by proxies.
x	Application specific identifier.

Apache web server maintains Common Log Format and Combined Log Format[3].

Common Log Format

```
LogFormat "%h %l %u %t \"%r\" %>s %b"
common
122.163.111.210 - - [22/Oct/2010:04:15:03
-0400] "GET
/imagesnew/misc_arrow_animated.gif
HTTP/1.1" 404 494
```

Combined Log Format

Another commonly used format string is called the Combined Log Format. It can be used as follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b
\"%{Referer}i\" \"%{User-agent}i\""
combined
```

This format is exactly the same as the Common Log Format, with the addition of two more fields. Each of the additional fields uses the percent-directive %{header}i, where header can be any HTTP request header. The access log under this format will look like:

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -
0700] "GET /apache_pb.gif HTTP/1.0" 200
2326 "http://www.example.com/start.html"
"Mozilla/4.08 [en] (Win98; I ;Nav)"
```

The additional fields are:

```
"http://www.silicon.ac.in/sitsbp/index.ht
ml"
```

```
(\"%{Referer}i\")
```

The "Referer" gives the site that the client reports having been referred from.

```
"Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; GTB0.0; SV1; .NET CLR
2.0.50727; .NET CLR 3.0.04506.30; .NET
CLR 3.0.4506.2152; .NET CLR 3.5.30729;
RediffIE8)"
```

```
(\"%{User-agent}i\")
```

The User-Agent HTTP request header. This is the identifying information that the client browser reports about itself.

More recent entries are appended to the end of the file. These data can be stored into a single file, or separated into distinct logs, such as an access log, error log, or referrer log.

Web usage mining research focuses on finding patterns of navigational behavior from users visiting website. These patterns of navigational behavior can be valuable when searching answers to questions like: How efficient is our website in delivering information? How the users perceive the structure of the website? Can we predict user's next visit? Can we make our site meeting user needs? Can we increase user satisfaction? Can we target specific groups of users and make web content personalized to them? Answer to these questions may come from the analysis of the data from log files stored in web servers. Web usage mining has then become a necessary task in order to provide web administrators with meaningful information about users and usage patterns for improving quality of web information and service performance. Successful websites may be those that are customized to meet user preferences both in the presentation of information and in relevance of the content that best fits the user.

IV. STEPS IN WEB USAGE MINING

A. Data Preparation

The information contained in a raw Web server log does not reliably represent a user session file [1]. 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: i) removing undesirable entries, ii) distinguishing among users, iii) building sessions, and iv) restoring the contents of a session.

1) Removing Undesirable Entries: Web logs contain user activity information, of which some is not closely relevant to usage mining and can be removed without noticeably affecting the mining, for example: a) All log image entries. The HTTP protocol requires a separate connection for every file that is requested 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. b) 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. Robot-access patterns can be identified from user-agent so many of the robot accesses can be detected and removed from the logs.

2) Distinguishing among Users: A user is defined as a single individual that accesses files from 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 behavior, users in the log must be distinguished. Figure 1 is a sample Web site where nodes are pages, edges are hyperlinks, and node "/index.php" 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.

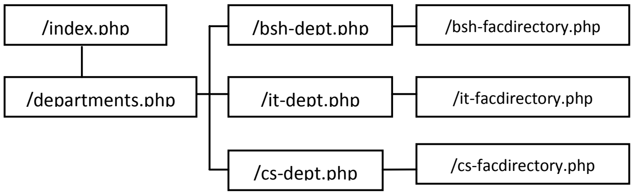

Figure 2: Sample website structure

The access data from an IP address (113.19.17.50) recorded on the log are given in Table 5. The paths are found by heuristics are index.php -- departments.php – bsh-dept.php – departments.php – cs-dept.php – cs-facdirectory.php – departments.php – bsh-dept.php – bsh-facdirectory.php – departments.php – it-dept.php -- it-facdirectory.php.

TABLE 5

SAMPLE LOG

clientIP	reqDateTime	URL	referer
113.19.17.50	24/Oct/2010:02:52:23 -0400	/index.php	-
113.19.17.50	24/Oct/2010:02:52:29 -0400	/departments.php	http://www.silicon.ac.in/index.php
113.19.17.50	24/Oct/2010:03:07:11 -0400	/bsh-dept.php	http://www.silicon.ac.in/departments.php
113.19.17.50	24/Oct/2010:03:07:33 -0400	/cs-dept.php	http://www.silicon.ac.in/departments.php
113.19.17.50	24/Oct/2010:03:08:21 -0400	/cs-facdirectory.php	http://www.silicon.ac.in/cs-dept.php
113.19.17.50	24/Oct/2010:03:08:44 -0400	/departments.php	http://www.silicon.ac.in/cs-facdirectory.php
113.19.17.50	24/Oct/2010:03:08:51 -0400	/bsh-dept.php	http://www.silicon.ac.in/departments.php
113.19.17.50	24/Oct/2010:03:08:55 -0400	/bsh-facdirectory.php	http://www.silicon.ac.in/bsh-dept.php
113.19.17.50	24/Oct/2010:03:09:18 -0400	/departments.php	http://www.silicon.ac.in/bsh-facdirectory.php
113.19.17.50	24/Oct/2010:03:09:21 -0400	/it-dept.php	http://www.silicon.ac.in/departments.php
113.19.17.50	24/Oct/2010:03:09:24 -0400	/it-facdirectory.php	http://www.silicon.ac.in/it-dept.php

3) Building Sessions (sessionization / episode identification): 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.

4) 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.

B. Pattern Discovery/ Pattern Extraction

It deals with extracting interesting patterns from the pre processed web logs. This is the key component of web usage mining. This can be subdivided into

Statistical Analysis:

TABLE 6
STATISTICAL ANALYSIS CATEGORY

Category	**Description**
General statistics	1) Total number of hits 2) Total number of visitors 3) Different errors 4) Successful visits 5) Incomplete visits 6) Error reports
Access statistics	Request Hit and Miss count based on 1) IP address 2) URL
Periodical statistics	Access of web pages according to period of time e.g. daily, monthly, yearly.

Clustering: clustering can be done according to i) Web pages, ii) Web page sequences, iii) client IP etc.

Classification: classify users according to their navigational behavior.

Association Rules: for example, thirty percent of department page viewers will enter the cs-dept pages.

V. CHALLENGES

The web usage mining algorithms are applied on the preprocessed web log data. The log files are collected from web server. But there are certain reasons due to which the actual logs are not collected. A) Due to the cache present on client browser, most of the request, if it is present in the cache are not sent to web server. B) Most of the time user does not visit the home page of a website. They directly navigate to a particular page, by getting the URL from search engines. So it reduces the hit count of index page. C) Generally in web pages designed by server side scripting like PHP, JSP or ASP.NET they use inner page. That is, one page consisting of more than one page. In that case the request for main page records two entries in access log. It is difficult to identify an inner page. D) Some web pages take query string as argument to the URL. E.g. *dept.php?dept=CSE*, *dept.php?dept=IT* like this. In this case the same page i.e. *dept.php* is accessed but with different arguments. It is difficult to count the page access of the web page without the argument.

In web usage mining the pattern extraction algorithms are applied on the log data after they are processed. So preprocessing is very much important and must be carried out with proper care. While preprocessing the web access log the above points should be taken into consideration so that it will produce a good set of access logs for pattern extraction.

VI. CONCLUSION

There is a growing trend among companies, organizations and individuals alike to gather information through web mining to utilize that information in their best interest. But it is a challenging task for them to fulfill the user needs and keep their attention in their website. Web usage mining has valuable uses to the marketing of businesses and a direct impact to the success of their promotional strategies and internet traffic. This information is gathered on a daily basis and continues to be analyzed consistently. Analysis of this pertinent information will help companies to develop promotions that are more effective, internet accessibility, inter-company communication and structure, and productive marketing skills through web usage mining. If we will be able to propose an efficient algorithm for the pattern extraction than it will help in the business of the website owners to understand their customer's behavior properly so that they can fulfill their requirements.

VI. REFERENCES

[1] Chen Hu, Xuli Zong, Chung-wei Lee and Jyh-haw Yeh, "World Wide Web Usage Mining Systems and Technologies", Journal of SYSTEMICS, CYBERNETICS AND INFORMATICS Vol. 1, No. 4, Pages53-59, 2003.

[2] FlorentMasseglia, Pascal Poncelet, Rosine Cicchetti, "An efficient algorithm for Web usage mining", Networking and Information Systems Journal. Volume X, 2000

[3] R. Pamnani, P. Chawan "Web Usage Mining: A Research Area in Web Mining"

[4] Qiankun Zhao, Sourav S. Bhowmick, "Sequential Pattern Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118 , 2003.

[5] S. Rawat, L. Rajamani, "Discovering Potential User Browsing Behaviors Using Custom-Built APRIORI Algorithm", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[6] Ming-Syan Chen, Jong Soo Park, Philip S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 10, NO. 2, MARCH/APRIL 1998.

[7] Jianhan Zhu, Jun Hong, John G. Hughes, "Using Markov Chains for Link Prediction in Adaptive Web Sites", Soft-Ware 2002, LNCS 2311, pp. 60–73, 2002

[8] WANG Tong, HE Pi-lian, "Web Log Mining by an Improved AprioriAll Algorithm", World Academy of Science, Engineering and Technology 4 2005

[9] Hengshan Wang, Cheng Yang, Hua Zeng, " Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", Communications of the IIMA 2006 Volume 6 Issue 2

[10] Paola Britos, Damián Martinelli, Hernán Merlino, Ramón García-Martínez, "Web Usage Mining Using Self Organized Maps", International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007

[11] Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman, "WEB USER NAVIGATION PATTERN MINING APPROACH BASED ON GRAPH PARTITIONING ALGORITHM", Journal of Theoretical and Applied Information Technology

[12] Kobra Etminani, Mohammad-R. Akbarzadeh-T., Noorali Raeeji Yanehsari, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", IFSA-EUSFLAT 2009

[13] Sandeep Singh Rawat, Lakshmi Rajamani, "DISCOVERING POTENTIAL USER BROWSING BEHAVIORS USING CUSTOM-BUILT APRIORI ALGORITHM", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[14] Mahdi Khosravi, Mohammad J. Tarokh, "Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method", 978-1-4244-8230-6/10/$26.00 ©2010 IEEE

[15] N. Sujatha, K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.464-476

[16] http://httpd.apache.org/docs/1.3/logs.html

[17] http://www.w3.org/TR/WD-logfile.html

[18] http://www.internetworldstats.com

[19] http://www.domaintools.com/internet-statistics/