Process of Web Usage Mining to find Interesting Patterns from Web Usage Data

Ketul B. Patel
A.M. Patel Institute of Computer Studies
Ganpat University
Kherva, India

Dr. A.R. Patel
Department of Computer Science
Hemchandracharya North Gujarat University
Patan, India

ABSTRACT

The traffic on World Wide Web is increasing rapidly and huge amount of data is generated due to users' numerous interactions with web sites. Web Usage Mining is the application of data mining techniques to discover the useful and interesting patterns from web usage data. It supports to know frequently accessed pages, predict user navigation, improve web site structure etc. In order to apply Web Usage Mining, various steps are performed. This paper discusses the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented Web Usage Mining applications and some Web Mining software.

Keywords

E-Commerce; Pattern Discovery; Web Mining; Web Server Log; Web Usage Mining

1. INTRODUCTION

Nowadays, the World Wide Web is growing continuously. Users interact frequently with different web sites and can access plenty of information on WWW. Web Mining is the application of data mining techniques to extract and analyze useful information from Web data [6]. Based on kind of data to be mined Web Mining can be classified into three different categories: Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining is the discovery of useful information from the contents of web documents. The Web document usually contains different types of data such as text, image, audio, video etc. Web Structure Mining focuses on analyzing the physical link structure of websites. The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. Web Usage Mining is the discovery of the activities of the users while they are browsing or navigating through Web. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site [9]. Knowledge discovered by Web Usage Mining, is useful in analyzing how the web pages are accessed or what are seeking for by the users and also to find weaknesses in the website structure.

The web usage data consists of the data from web server logs, browser logs, proxy server logs and user profiles. In Web Usage Mining, data mining techniques are applied to pre-processed web log data in order to find interesting and useful patterns. Visitors' browsing behaviour is recorded into web server log. The questions can be answered by analysing log files such as what pages are being accessed frequently? From what search engine are visitors coming? Which browser and operating systems are most commonly used by visitors?

E-commerce has provided effective way of doing business by electronic transactions through internet. For Web Usage Mining in e-commerce, the data of customer profile, inventory and demographic information from other relational databases are integrated with web usage data and visitors' behaviour patterns can be discovered by applying data mining techniques such as Association Rules, Sequential Analysis, Clustering and Classification. Web Usage Mining can help e-commerce companies to improve the web site, attract visitors, to provide personalized and adaptive service to regular user, identify potential customers for e-commerce, supporting business intelligence and marketing decisions etc.

2. THE PROCESS OF WEB USAGE MINING

The Web Usage Mining is the application of data mining technique to discover the useful patterns from web usage data. It can discover the user access patterns by mining log files and associated data of particular web site. Figure 1 shows the process of Web Usage Mining consisting steps Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis.

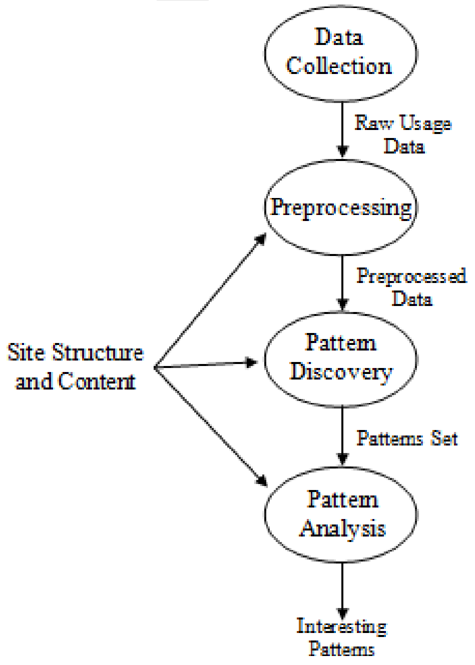


Fig 1: The Process of Web Usage Mining

2.1 Data Collection

The web log files on the web server are major source of data for Web Usage Mining. When user requests resources of web server, each request is recorded in the web log file on web server. As a consequence, users browsing behavior is recorded into the web log file. In Web Usage Mining, data can be collected from server log files that include web server access logs and application server logs. Data is also obtained from site files and operational databases.

2.2 Pre-processing

The data collected in web log file is incomplete and not suitable for mining directly. Pre-processing is necessary to convert the data into suitable form for pattern discovery. Pre-processing can provide accurate, concise data for data mining. Data pre-processing, includes data cleaning, user identification, user sessions identification, path completion and data integration [1].

2.3 Pattern Discovery

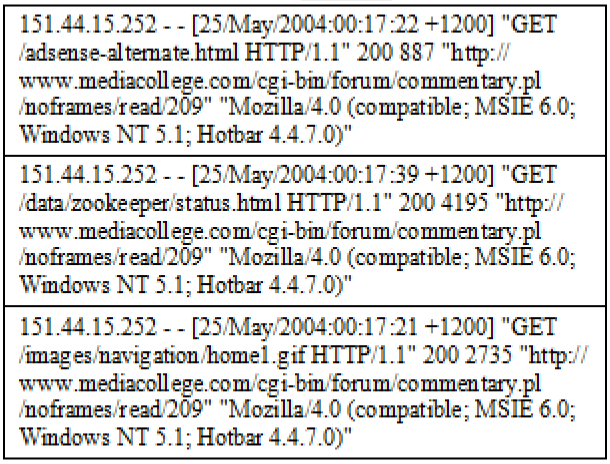
To discover the novel, potentially useful and interesting information, several methods and data mining algorithms are applied such as Path Analysis, Association Rules, Sequential Patterns, Clustering, Classification etc.

2.4 Pattern Analysis

Pattern analysis techniques are used to highlight overall patterns in data and to filter out uninteresting patterns. The techniques like Knowledge Query Mechanism, OLAP and visualization are used for pattern analysis.

3. WEB SERVER LOG

When user navigates the web pages of the website, the browser sends requests to the website server for resources accessed by the user. Each request is recorded by the server in so-called access log files or server log files. The log files keep a lot of information about each user's access to the web server. Each line or entry in the web log represents a request for a resource. The different web server support different log format. The exact content of an entry varies from log format to log format. Nearly all the server log file contains information such as Visitor's IP address, access date and time, URL of page accessed, the status code returned by the server, bytes transferred to the user etc. Most of log files of web servers are stored in a common log file format or in an extended log file format. Figure 2 represents the sample web server log entries [12].



```
151.44.15.252 - - [25/May/2004:00:17:22 +1200] "GET
/adsense-alternate.html HTTP/1.1" 200 887 "http://
www.mediacollege.com/cgi-bin/forum/commentary.pl
/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; Hotbar 4.4.7.0)"

151.44.15.252 - - [25/May/2004:00:17:39 +1200] "GET
/data/zookeeper/status.html HTTP/1.1" 200 4195 "http://
www.mediacollege.com/cgi-bin/forum/commentary.pl
/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; Hotbar 4.4.7.0)"

151.44.15.252 - - [25/May/2004:00:17:21 +1200] "GET
/images/navigation/home1.gif HTTP/1.1" 200 2735 "http://
www.mediacollege.com/cgi-bin/forum/commentary.pl
/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; Hotbar 4.4.7.0)"
```

Fig 2: Sample Web Server Log Entries

Web servers can be configured to write different fields into the log file in different formats. The most common fields used by web servers are the followings: IP Address, Login Name, User Name, Timestamp, Request, Status, Bytes, Referrer, User agent. The explanation of fields included in sample web log entries of Fig. 2 is as under.

1) *IP Address:* IP address of the remote host that made the request. If DNS host name is available then it will be substituted in place of IP address.

2) *Login Name:* Login name of the user making the HTTP request. If there is no value, a dash (-) is substituted instead of the login name.

3) *User Name:* The name of the authenticated user who access the server by making the HTTP request, anonymous users are represented by a dash (–)

4) *Timestamp:* Time-stamp indicates date and time of the visit with the web server's time zone.

5) *Request:* It indicates HTTP request. It consists of three pieces of information: the HTTP method, the requested resource and HTTP protocol version.

6) *Status Code:* The success or failure of the http request is indicated by HTTP status code. For example, "200" is success, "404" for File Not Found, "500" for Internal Server Error.

7) *Bytes transferred:* It indicates number of bytes transferred to the user as part of the HTTP request.

8) *Referrer:* The URL of the page the visitor was on when he clicked to come to current page.

9) *User agent:* It represents the software used by the visitor to access the site. It could be a web browser, web robot or link checker etc. It also indicates the operating system used by the client. MSIE 7.0 means Internet Explorer 7. Windows NT 6.0 indicates Windows Vista. Windows NT6.1 indicates Windows 7.

A visitor request for web page can generate different web log records. The first record documents the request of user to particular web page and remaining records are created for requests made to obtain images on the requested page. The quantity of information stored in the log files is very large. The raw web log data is usually diverse and incomplete and difficult to be used directly for further pattern mining. The web log data can be preprocessed in order to obtain session information for all users and data reduction techniques can be applied to web log data in order to obtain raw sets for data mining. Before it can be used for anything at all, data must be extracted from the log file and analyzed.

4. DATA PRE-PROCESSING

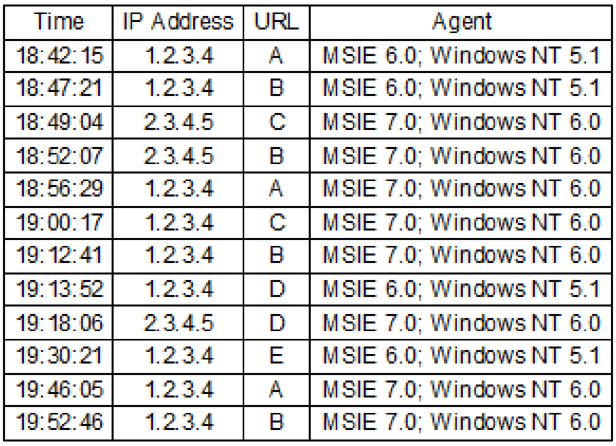
The data collected from web server log is often incomplete and creates uncertainty. The data pre-processing is necessary in order to clean, correct and complete input data and to mine the knowledge effectively. Pre-processing can provide accurate and concise data for data mining. Data pre-processing task includes data cleaning, user identification, user sessions identification, path completion and data integration [3]. In data pre-processing the server sessions are identified which represents the information such as who accessed the website, what pages were requested and how long each page was viewed.

4.1 Cleaning

The items which are not related for usage analysis must be removed from the log files. When user requests to particular page from web server, various log entries are recorded. If page contains the images, videos, scripts, flash animations etc then resource requests for them will also be added in the log file. The objective of Web Usage Mining is to find users' behaviour. So the entries for these resource requests do not make sense and must be removed from log file. Elimination of irrelevant items can be done by checking the suffix of the URL, which signifies in what format the kinds of files are. For example, the entries from log file with URL suffix jpg, gif, css, js, mov, avi, swf etc can be removed [10]. The list of suffixes can be changed depending on the type of site being analysed. It also involves removing requests where the server returned an error code. Erroneous log records are generated when access requests fail for various reasons and can be identified by HTTP status code. The Resulting status code may have different value, for example "200" is success, "404" File Not Found, "500" Internal Server Error. The entry can be removed if server status code is not "200" something. We can consult at RFC 2616 for complete list [8]. In this way other types of requests can also be eliminated. Cleaning the data significantly reduces the number of log entries we have to work with and it is essential for finding usable access patterns.

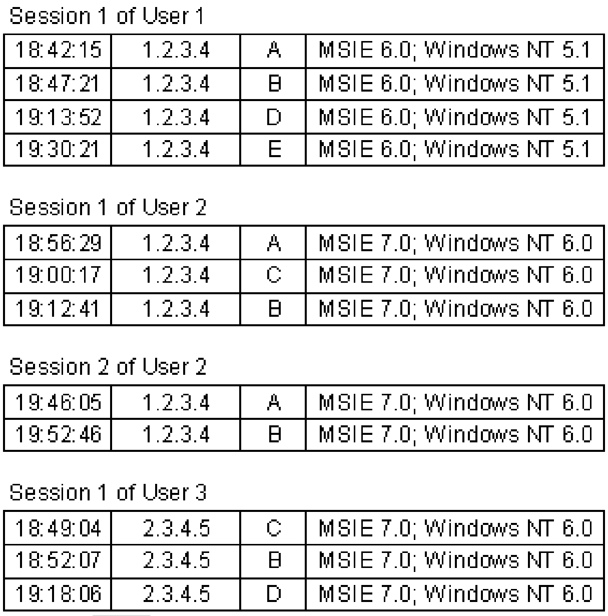
4.2 User Identification and Session Generation

After data cleaning, unique users must be identified. To identify the users, one simple method is to use login information, if users log in before using the web-site or system. Another approach is to use cookies for identifying the visitors of a web-site by storing a unique ID. However, these two methods are not general enough because they depend on the application domain and the quality of the source data. We can use a more general method to identify user. A new IP indicates a new user. The same IP but different user agent means a new user. The user agents are said to be different if it represents different web browsers or operating systems in terms of type and version. The list of log entries is sorted by the combination of IP addresses or host name of the user and the user agent [11]. The result is a list where all entries generated by the same user are clustered together and stored as separate log entry lists.



Time	IP Address	URL	Agent
18:42:15	1.2.3.4	A	MSIE 6.0; Windows NT 5.1
18:47:21	1.2.3.4	B	MSIE 6.0; Windows NT 5.1
18:49:04	2.3.4.5	C	MSIE 7.0; Windows NT 6.0
18:52:07	2.3.4.5	B	MSIE 7.0; Windows NT 6.0
18:56:29	1.2.3.4	A	MSIE 7.0; Windows NT 6.0
19:00:17	1.2.3.4	C	MSIE 7.0; Windows NT 6.0
19:12:41	1.2.3.4	B	MSIE 7.0; Windows NT 6.0
19:13:52	1.2.3.4	D	MSIE 6.0; Windows NT 5.1
19:18:06	2.3.4.5	D	MSIE 7.0; Windows NT 6.0
19:30:21	1.2.3.4	E	MSIE 6.0; Windows NT 5.1
19:46:05	1.2.3.4	A	MSIE 7.0; Windows NT 6.0
19:52:46	1.2.3.4	B	MSIE 7.0; Windows NT 6.0

Fig 3: A List of Sample Log Entries



Session 1 of User 1

18:42:15	1.2.3.4	A	MSIE 6.0; Windows NT 5.1
18:47:21	1.2.3.4	B	MSIE 6.0; Windows NT 5.1
19:13:52	1.2.3.4	D	MSIE 6.0; Windows NT 5.1
19:30:21	1.2.3.4	E	MSIE 6.0; Windows NT 5.1

Session 1 of User 2

18:56:29	1.2.3.4	A	MSIE 7.0; Windows NT 6.0
19:00:17	1.2.3.4	C	MSIE 7.0; Windows NT 6.0
19:12:41	1.2.3.4	B	MSIE 7.0; Windows NT 6.0

Session 2 of User 2

19:46:05	1.2.3.4	A	MSIE 7.0; Windows NT 6.0
19:52:46	1.2.3.4	B	MSIE 7.0; Windows NT 6.0

Session 1 of User 3

18:49:04	2.3.4.5	C	MSIE 7.0; Windows NT 6.0
18:52:07	2.3.4.5	B	MSIE 7.0; Windows NT 6.0
19:18:06	2.3.4.5	D	MSIE 7.0; Windows NT 6.0

Fig 4: Generated Sessions

Identification of the user sessions is very important because it constitutes a basic processing unit for discovery of interesting, prominent access patterns and so it largely affects the quality of pattern discovery result. The construction of the user activities in sessions for a particular Web site contains a correct mapping of activities to distinct web users. A long sequence of visits by the users has to be broken into user sessions. A user session is a set of pages visited by the same user within the duration of one particular visit to a web site. Each user session consists of only the pages visited by a user in a row. We consider web log data as a sequence of distinct web pages, where subsequence, such as user sessions can be observed by unusually long gaps between consecutive requests. A session is a collection of log entries generated by the page navigation of a single user within a certain time frame.

User sessions are identified on the basis of the user identification. Because of the stateless connections between the client and the web server, Web log records have no visit session designations. Therefore, to reconstruct visit sessions from Web log records, we use the common timeout method, according to which a visit session is considered terminated when the time elapsed between page requests exceeds a specified limit. A set of pages visited by a specific user is considered as a single user session if the pages are requested at a time interval not larger than a specified time period. Generally a time period of between 20 and 30 minutes is considered optimal. When the time interval exceeds the specified time it takes a person to read a web page or the user has left the browser to do other things. The timeout method is fairly effective and accurate for identifying visit sessions. According to Spiliopoulou, more than 90% of genuine visit sessions can be identified by the timeout method [3]. After generating the sessions by sorting the log entries on IP addresses and user agent, we check whether any sessions can be further split up by checking how much time has passed between two consecutive log entries. If the time is above a specific threshold value, the session is split into two sessions. Figure 3 shows a list of sample log entries and Figure 4 shows sessions generated from it by considering timeout 30 minutes. The whole session set generated are used as raw data for pattern discovery phase.

4.3 Path Completion

Due to existence of page cashing technology and proxy servers, some important accesses are not recorded into web server log. It results into missing access references to those pages that have been cashed. Users' access patterns are not reflected accurately by incomplete access log. Path completion step is carried out to acquire the missing reference. The referrer information in server logs and knowledge of site topology can be used for effective path completion.

If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. If the page is in the user's recent request history, the assumption is that the user backtracked with the back button available on most browsers, calling up cached versions of the pages until a new page was requested. The site topology can also be used for the same, if referrer information is not clear. If more than one page in the user's history contains a link to the requested page, the page closest to the previously requested page is assumed as a source of the new request [11]. In this way, the missing page references are found and added to the user sessions.

4.4 Data Integration

In e-commerce, the operational databases contain user data, product, purchases etc. The user data provides demographic information about users and consists of registration and customer profile information. These data must be integrated with pre-processed data while applying Web Usage Mining in e-commerce in order to improve business intelligence [13].

5. PATTERN DISCOVERY

After identifying user sessions, the various techniques of web usage pattern discovery are applied in order to detect interesting and useful patterns. There are several kinds of access pattern mining that can be performed depending on the needs of the analyst. Some of pattern discovery techniques are discussed below.

5.1 Path Analysis

A graph can be formed to perform path analysis. A graph represents some relation defined on web pages. The physical layout of the web site can be represented by graph in that web pages are nodes and link between pages are directed edges. Using it frequent paths traversed by users, entry and exit points can be determined easily [4]. For example what paths do users traverse before they go to particular URL? What percentage of clients left the site after five or less page references?

5.2 Association Rules

The correlations between web pages that are most often referenced together in a single user session can be discovered by association rules. In e-commerce, association rules can be used to find the relevant pages browsed by customers. It can provide the information: What are the set of pages frequently accessed together by web users? What page will be fetched next? What are paths frequently accessed by web users?. Implement association rules to on-line shopper can generally find out his/her spending habits on some related products [2]. For example, if a transaction of an on-line shopper consists of a set of items, while each item has a separate URL. Then the shopper's buying pattern will be recorded in the log file, and the knowledge mined from it, can be the form like this: x% of clients who accessed the web page with URL A.html, also accessed B.html, y% of clients who accessed S.html, placed an online order in P.html. The association rule can be used to restructure the web site by adding links that interconnect pages which are often viewed together.

5.3 Sequential Patterns

The technique of sequential pattern discovery can be applied to web server logs. It attempts to find intersession patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions [14]. In e-commerce, customer access data is recorded in the web server log with a unit of a time period. Using sequential pattern discovery, useful user trends can be discovered, website navigation can be improved and adopt web site contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles. The user's visit patterns can also be predicted using sequential pattern discovery in web server log and helps in targeting adverting aimed at groups of users based on these patterns. The sequential patterns can be discovered as the following form: x% of client who bought items using url A, also placed an order online within 10 days using url B.

5.4 Clustering

Clustering can identify user or data items with common characteristics. A group of users can be clustered that have similar navigation patterns on a web site. In e-commerce using cluster analysis technique, a group of customers can be clustered with similar browsing behaviour and common characteristics of customers can be analysed. These can help the e-commerce users to get better understanding to their customers and customer-oriented service can be provided. Cluster analysis can help with marketing decisions [7].

Clustering of pages will discover groups of pages having related content. An example of clustering could be: X % of clients who bought gold coin using url A, were in the 35-40 age group, with annual income between $50,000 – 60,000.

5.5 Classification

In classification analysis, data items are classified according to predefined categories. If it is needed to develop a profile of user belonging to a particular class or category then features are extracted that well describe the properties of given class or category. There are many algorithms such as decision trees, neural networks, Bayesian classifier, and probability theory for classification [4].

After classification, business activities can start according to the characteristics of this type of clients, providing targeted, personalized information services. In e-commerce, after classifying the data with the same characteristics, e-commerce enterprise can provide personalized information services according to the characteristics of such customers.

6. PATTERN ANALYSIS

The final step in Web Usage Mining process is Pattern analysis. Using pattern discovery, the pattern set is found and then pattern analysis is performed to select the interesting patterns and to filter out uninteresting patterns. The patterns are analysed using techniques such as Data & Knowledge Querying, OLAP techniques and Usability analysis. Data & Knowledge querying mechanism uses a tool like SQL. In OLAP techniques, the result of pattern discovery is loaded into data cube and then OLAP operations are performed. After this, to interpret the results, visualization techniques are used [4]. The result of pattern analysis helps to improve the system performance and to modify the web site. It helps to attract the visitors and to give the personalized services to regular user.

7. WEB USAGE MINING APPLICATIONS

In E-Commerce, Web Usage Mining provides useful information that can help to improve customer, sales and marketing support. Some applications of Web Usage Mining are as below.

- Improving the design of e-commerce web site according to user's browsing behavior on site in order to better serve the needs of users.
- Personalizing web sites according to individual's interest and make dynamically changing particular web site for visitor.
- Developing a security system that can detect the intrusion and to restrict the user's access to certain online contents.
- Understanding customers' need and retaining them by providing customized products, improving satisfaction with help of tracking browsing behavior in e-commerce.
- Evaluating the effectiveness of advertising by analyzing large number of consumer behavior patterns.

8. WEB MINING SOFTWARE

There are commercial and free open source Web Mining software available. Some of them have been listed.

- WebLog Expert
- Click Tracks
- 123LogAnalyzer
- Amadea Web Mining
- Nihuo Web Log Analyzer
- Megaputer WebAnalyst
- AlterWind Log Analyzer Professional
- Conversion Track from Antssoft
- Surf Pattern Visual Analyzer
- ANGOSS KnowledgeWebMiner
- Analog
- jwanalytics
- htminer
- Visitator
- WUM : Web Utilization Miner [5]

9. CONCLUSION

Due to large amount of information on WWW, it is a rich area of Web Mining. Many users interact with web sites daily. Web Usage Mining is a technique of data mining to extract interesting and valuable information from web usage data.

In this paper, we have described Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis tasks of Web Usage Mining process. Pre-processing is done to make the data available in appropriate form for Pattern Discovery. Due to existence of local cache and proxy server, some page references are not recorded into web server log. To deal with this issue, referral log and knowledge of site topology is used to find out missing reference. Once the user sessions are identified, useful patterns are discovered by applying data mining algorithms such as Association Rules, Sequential Patterns, Clustering, Classification etc. Pattern analysis helps to filter out uninteresting patterns from pattern set. Web Usage Mining can help e-commerce companies for understanding customer behaviour, optimizing web site, improving customer services and relationship, measuring the effectiveness of marketing efforts, providing personalized services to customers.

10. REFERENCES

[1] Li Mei, Feng Cheng, "Overview of Web Mining Technology and Its Application in E-commerce", 2nd International Conference on Computer Engineering and Technology, 2010, Volume 7, pages 277-280

[2] Penelope Markellou, Ioanna Mousourouli, Spiros Sirmakessis, Athanasios Tsakalidis, "Personalized E-Commerce Recommendations", Proceedings of IEEE International Conference on e-Business Engineering, 2005.

[3] Spiliopoulou M., Mobasher B., Berendt B., and Nakagawa M. "A framework for the evaluation of session reconstruction heuristics in web-usage analysis", INFORMS Journal on Computing, 15, 2003, pages 171-190.

[4] Jian-Guo Liu, Wei-Ping Wu, "Web Usage Mining for Electronic Business Applications ", Proceedings of the Third International Conference On Machine Leaming and Cyhemetics, Shangha, August 2004, 26-29.

[5] Sheilini Jindal, Gaurav Kumar, "A Proportional Analysis on the Illustrious Practices for the Extraction and Discovery of Hidden Patterns - Data and Web Mining", International Journal of Enterprise Computing and Business Systems, Vol. 1, Issue 1, January 2011.

[6] G. Chang, M. J. Healy, J. A. M. McHugh, and J. T. L. Wang, "Mining the World Wide Web: An Information Search Approach", Kluwer Academic Publishers, 2001.

[7] Sung-Shun Weng, Mei-Ju Liu, "Personalized product recommendation in e-commerce", IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004.

[8] RFC 2616 - Hypertext Transfer Protocol - HTTP/1.1. http:// www.faqs.org/rfcs/rfc2616.html. Vigente al 19/11/2005.

[9] Y. Fu, M. Shih, "A Framework for Personal Web Usage Mining, International Conference on Internet Computing", Las Vegas, NV, pages 595-600.

[10] Lalani, A.S., "Data mining of web access logs", School of Computer Science and Information Technology. Royal Melbourne Institute of Technology. Melbourne, Victoria, Australia, 2003.

[11] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", http://maya.cs.depaul.edu/mobasher/papers/webminer-kais.pdf

[12] "Server Log Files", http://www.mediacollege.com/internet/statistics/ logs/

[13] Bamshad Mobasher, "Web Usage Mining", http://maya.cs.depaul.edu/~mobasher/papers/12-web-usage-mining.pdf

[14] Manoj Pandia, Subhendu Kumar Pani, Sanjay Kumar Padhi, Lingaraj Panigrahy, R.Ramakrishna, "A Review Of Trends In Research On Web Mining", International Journal of Instrumentation, Control & Automation, Volume 1, Issue 1, 2011, pages 37-41.