“An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining”

Brijesh Bakariya[1], Krishna K. Mohbey[2] G. S. Thakur[3]
Department of Computer Applications
M.A.N.I.T., Bhopal-462051
brijesh_scs@yahoo.co.in[1]
kmohbey@gmail.com[2]
ghanshyamthakur@gmail.com[3]

Abstract: Several data mining techniques applied in Web usage mining applications for discovering user access pattern from web log data. To understand and provide better services it will require Web-based applications. Web usage mining is one of the types of Web mining. Web mining is the technique to extract knowledge from web content, structure and usage. It is the collection of technologies to accomplish the possible of extracting valuable knowledge from the World Wide Web and its usage pattern. Web mining enables to find out relevant result from Web data including web document, hyperlink between documents, usage log of website etc. There are three main areas of web mining research –content, structure and usage. This paper provide an overview of previous and existing work in all three areas, and also define an overview of data preprocessing process like Data Cleaning, User Identification, Session Identification, Transaction Identification, Path Completion used in Web usage mining.

Keywords: data mining, web content mining, web structure mining, web usage mining, data preprocessing.

1 Introduction

Today is the day of Information Technology, accessing information is the most frequent task. Day by day we have to go through various kind of information that we require, we have to just browse the web and get desired information with a single click. Now a day, internet is playing such a crucial role in our daily life that is very difficult to survive without it, because millions of electronic data are included on hundreds of millions data that are previously online today. The amount of data on World Wide Web are huge therefore it is very critical to store all data in an organized way, it also produced problem in data accessing.

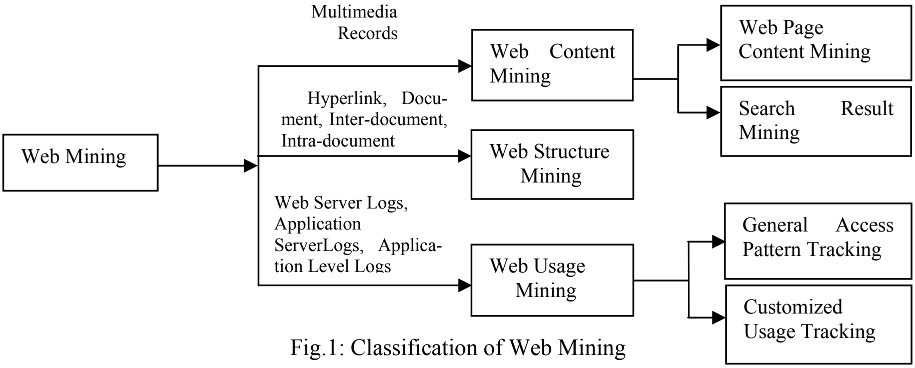

Multimedia Records
Web Content Mining
Web Page Content Mining
Search Result Mining
Hyperlink, Document, Inter-document, Intra-document
Web Mining
Web Structure Mining
Web Server Logs, Application ServerLogs, Application Level Logs
General Access Pattern Tracking
Web Usage Mining
Customized Usage Tracking

Fig.1: Classification of Web Mining

J. C. Bansal et al. (eds.), *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012),* Advances in Intelligent Systems and Computing 202, DOI: 10.1007/978-81-322-1041-2_35, © Springer India 2013

The World Wide Web has affected a lot to both users as well as the web site owners. The web site owners are able to achieve to all the targeted viewers countrywide and globally [27]. To extract frequent data from huge collection of data [11], data mining techniques can be applied. But the web data is unstructured and semi-structured, so we can not directly apply the technique of data mining .To a certain extent another application is evolved called web mining [10], which is applied on web data. There are several problems like improving web sites, to better understand the visitants behaviour, e-commerce, e-business, advertisements with the help of web mining we can discover interesting patterns in all above problems. Web usage mining [9] is accomplished first by coverage visitors transfer information, which is based on Web server log files and other source of transfer data .Web server log files were used primarily by the webmasters. They may be web architect, web developer, site author, or website administrator and system administrators like Databases, configuring a computer systems, software etc. Web server log files are used to contain the details of the user behaviour [14] and transactional details. These also stores the overall activities of the all users who access the websites.log files may contain the time of session start, details of web page which is access by user, traffic details or error information etc. Web log files plays vital role in web mining process because they provide overall details to the administrator for improving the performance of website in the World Wide Web environment. Web traffic data are handling through Web log file [13] this is the one way. Another way is to find out TCP/IP packets as they cross the network, and to attach to each Web server. After the Web traffic data is obtained, it may joint with other relational databases, over which the data mining techniques are implemented. By different data mining techniques such as association rules mining, path analysis, sequential analysis, clustering and classification, using all these techniques visitors' behaviour patterns are found and interpreted.

1.1 Web Content Mining

It refers to the extraction of useful information from huge data according to the contents. It obtains data from the web pages according to given contents. These contents can either text based or multimedia based. Web content mining generally deal with documents in text or html format and it also get information on the bases of image, audio, video or other contents [28].

1.2 Web Structure Mining

This mining refers to the process of obtaining required information from the structural patterns. It gets information from the websites on the bases of the links and documents relationships [29]. For this mining process websites documents are generally arranged in the tree structure which describes the relationship between different documents. When a user try to search a particular page on the web, similar pages also reflects on the results [28].

1.3 Web Usage Mining

Web usage mining extracts useful information by using the server logs. Server logs stores the accessing patterns of the user in the form of URL, IP addresses or visiting times etc. by these log data, one can collect the behavioural patterns of the users. These patterns are used to define pattern discovery and associations between documents [28] [29].

2 Web Usage Mining Process

2.1 Data Pre-processing

There are lots of issues for pre-processing like Data Collection, Data Integration, and Transaction Identification. Pre-processing is a methods for converting the Content information, Structure information and usage information enclosed in the different presented data sources into the data abstractions necessary for pattern discovery. Log file pre-processing [18] consists of data cleansing, user identification, session identification. In data cleansing irrelevant records are eliminated. Records with GIF, JPEG, and CSS and so on as suffixes are eliminated. In the second step, we have the task of user and session identification is to find out the diverse user sessions from the original web access log. One way is to branching them based on their IP addresses.

2.2 Pattern Discovery

It is a process to find out patterns in web logs but is frequently approved only on samples of data. The mining process will be unsuccessful if the samples are not a good representation of the larger body of data [19]. According to Literature reviews following methods are used for pattern discovery process:

A. Statistical Analysis
B. Association Rules
C. Clustering
D. Classification
E. Sequential Patterns
F. Dependency Modeling

2.2.1 Statistical Analysis

It is the most general method to take out knowledge about visitors to a web site. We can perform different kinds of expressive statistical analyses like mean, median, mode, frequency etc [33]. On variables such as page visit, the time of visit and navigational path length. There are various web traffic analysis tools produce which generate an intervallic report containing statistical information such as the most commonly accessed pages, average view time of a page or length of navigational path.

2.2.2 Association Rules

It is a procedure for finding frequent patterns, correlations and associations [31] among sets of stuffs and it is used to relate pages that are most frequently located together in a single server session. Association rules [23], [26] are used in order to disclose correlations among pages accessed together throughout a server session. Those types of rules point out the possible relationship between pages that are often viewed together even if they are not directly connected, and can disclose associations between groups of users with specific interests.

2.2.3 Clustering

Clustering is used to group together a set of items that have similar characteristics. In the Web Usage Mining, there are two kinds of interesting clusters to be discovered user clusters and page clusters [20]. User clustering results in groups of users that seem to behave similarly when navigating through a Web site and Page clustering identifies groups of pages that appear to be conceptually related according to the user's perception.

2.2.4 Classification

Classification is the process of mapping a data into one of several predefined classes [6]. In the Web area, one is interested in developing a users profile belonging to a particular category or class. These necessitate selection and extraction of features that best explain the properties of a known class or category. Classification can be done by using supervised inductive learning algorithms [25] such as k-nearest neighbor classifiers, Vector Machines, decision tree classifiers, naive Bayesian classifiers etc.

2.2.5 Sequential Patterns

Sequential patterns indicate the correlation between transactions [32]. The method of sequential pattern discovery challenged to find inter-session patterns such that the presence of a set of objects is followed by another object in a time-ordered set of episodes or session. With the help of this approach, Web marketers can forecast future visit patterns which will be helpful in placing advertisements intended at certain user groups.

2.2.6 Dependency Modeling

The aim to develop this method is to prepare a model which is capable of representing significant dependencies between various variables in web domain. There are different kinds of learning techniques such as Bayesian Belief Networks and Hidden Markov models which can be employed to model the browsing behaviour of users. These models are also useful to analyse the behaviour of the users. Modelling of web usage patterns will not be provide a theoretical framework of users but is useful in forecasting future web resource utilization. By these models, future web resource consumption can be predicted [33].

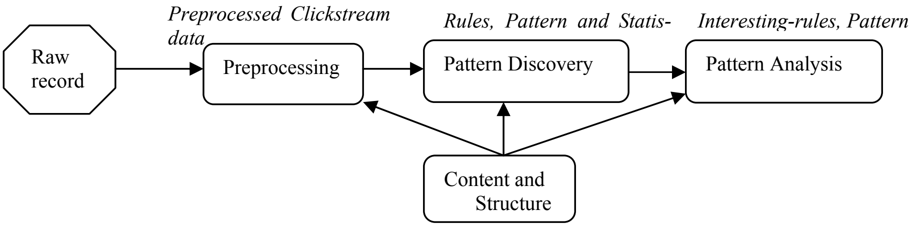

Preprocessed Clickstream data
Rules, Pattern and Statis-
Interesting-rules, Pattern
Raw record
Preprocessing
Pattern Discovery
Pattern Analysis
Content and Structure

Fig2: Web Usage Mining Process Model

3 Literature Review

Mohd Helmy et al. [1] describes the pre-processing techniques on IIS Web Server Logs ranging from the raw log file until before mining process can be performed. Pre processing is very important and essential for data mining process, pre processing activities can be applied in various ways; it depends on the purpose of algorithm and nature of the applications. Ms. Dipa Dixit et al. [2] discuss two different approaches for data preprocessing one based on XML and other based on text file. But the way and steps involved in pre-processing are considered same for both the approaches. Arshi Shamsi et al. [3] presents, how web server log data is preprocesses, which includes data cleaning, user identification and Sessionization, path completion. If the data is preprocessed by some techniques it is used for discovering some useful patterns. T. Revathi et al. [4] describes an efficient approach for data pre-processing for mining Web based user data in order to speed up the data preparation process. It provides flexibility for data preprocessing and reduce complexity and difficulty of preparation for mining user data. However, we can't directly performed data mining process directly on the Web log data because of the messy and redundant content and other reasons. This paper describes the data pre- processing techniques for Web log data in order to meet the needs of data mining. M. Malarvizhi et al. [5] identifies the problems in existing techniques of preprocessing. It also proposes the possibility of improving the performance of preprocessing with several experiments. The experimental results show that the log error rate, log sizes are reduced and the quality is improved. Suneetha K.R et al. [6] presents algorithm for data cleaning, user identification and session identification. The main new approach of this paper is to access the usage pattern of preprocessed data using snow flake schema for easy retrieval.

4 Web Log Data and its Attributes

Web log file is log file that automatically created and maintained by web server. Every click on the website, include the HTML document, images or other objects are logged. It is essential that every raw web file format on one line of text for each click on the website. This contains information about the users who have already visited the sites. More recent entries are complicated to append at the end of the file. There are various attributes in log files [24] which are mentioned in the table1. This is the statistical analysis of server log which have used to examine traffic pattern on the time of the day, day of week, or user agent. Efficient web site administration adequate hosting resources and the fine turned off sales efforts can be aided by analysis of the web server logs.

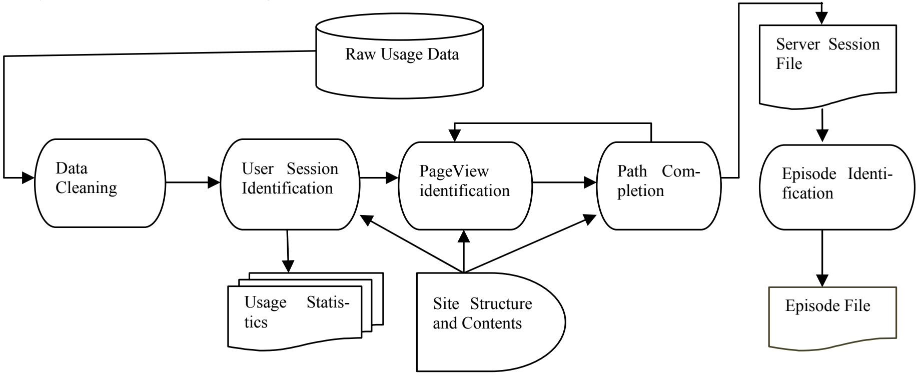


Fig.3: Preprocessing Model for Web Data

The following are the data preprocessing steps-

5.1 Data Cleaning

It is a process in which noise, unused and irrelevant data are removed [11]. It is also useful for web usage mining to clean [15] server log and to eliminate irrelevant information are of importance for any type of web log analysis. The discovery of associations or statistical report are only useful if data represented in the server log and it also gives the accurate picture of the user access to the web site, it is extremely significant, because only this log data that is able to accurately reflect the patterns of user access can be useful to search the correctness of the knowledge, get the model and the results meaningful. In web server log the problem arise when the HTTP protocol requires a separate connection for every file that is requested from the web server. When a user download a particular page then there are different elements are also downloaded with pages like graphics and scripts. In server log entries these all element details are stored. In most cases, only the log entry of the HTML file request is relevant and should be kept for the user session file then the Solution for that problem is to Eliminate some items deemed irrelevant can be reasonably accomplished by checking the suffix of URL name. All log entries with file name suffixes such as gif, jpeg etc. so that the list can be changed according to the site being analysed [33].

Table 1: Sample of web log

IP address	User Name	Timestamp	Access Request	Result Status Code	Bytes Transf-erred	Referrer URL	User Agent
123.456.78.2	U1	[25/Apr/1998:03:04:41 -0500]	GET XYZ.html HTTP/1.0	200	1923	XYZ.html	Mozilla/4.7[en]C-SYMPA (Win95; U)
123.456.78.9	U2	[25/Apr/1998:03:05:20 -0500]	GET PQR.html HTTP/1.0	200	2828	PQR.html	Mozilla/4.05 (Macintosh; I; PPC)
123.456.78.3	U3	[25/Apr/1998:03:06:20 -0500]	GET ABC.html HTTP/1.0	200	952	ABC.html	Mozilla/4.05 (Macintosh; I; PPC)

5.2 User Identification

User Identification Process comes after the log file has been cleaned. User Identification [16] means recognizing the user. It is the key part of the process of the server session identification. The identification of users is a very difficult task because of local caches and proxy servers.

5.2.1 User identification by IP address

IP address (computer address) is unique address for each user while browsing the website. So we can just consider that every new IP address represents a new user. But, it is poor user identification method when we are using the user's IP, because of the following problems:

1. Several users can be used the same IP address or computer (i.e. college, internet cafe etc.), so we do not know how many users hidden behind one IP address.
2. One user can have different IP addresses, since a user accesses the Web from different machines will have different IP address.
3. One user can use multiple browsers for the same IP address.

5.2.2 User identification using User registration Data:

Mostly website uses username and password for user identification. When user want to login a website; username and password are essential. These entries are also stored in the web log files; and useful for the next login. But these facilities are not available in every website so that it is not appropriated for the general web browsing [16].

5.2.3 User identification using Cookies:

Cookies are used to store temporary data while WebPages are downloaded on the client, it provides fast accessing if the request come for same data again. They are helpful to solve the problem of user identification. Cookies are HTTP headers in string format. By using Cookies we can extract the details of users and resources which are accessed by the user. If cookies are used for user identification then two problems can be arrived; first if the user lock the use of cookies the server can't store data on local machine. Second, user can delete the cookies. Therefore this technique is not reliable always. [9]

5.3 Session Identification

Session identification process comes after the user identification process. In this process we identify the session of users [17]. If a particular user visited the same site more than one then log entries can be divided in sessions [12].

In other word, if we group the different activities of a single user in the web log files is called session [30]. When a new user starts web page browsing, a new session is created, mostly sites define the time duration of session. Within this session duration that user can visit on multiple pages and these transactions are stored in log files.

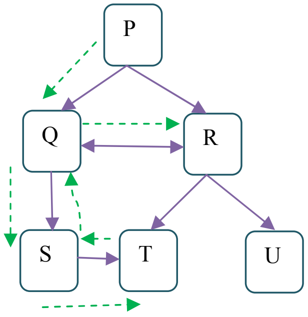
Timeout is one of the methods for session identification; it uses assumption for the time duration between two page requests. If this predefines time exceeded then new session is started automatically [7].If proxy servers [8] are uses then log files are creates problems for session identification.

5.4 Transaction Identification

The goal of transaction identification is to create meaningful group of references for each user. By considering the time consumed by users in viewing the page, pages can be categorized as auxiliary or content page. Auxiliary pages are used to navigate from one page to another. Otherwise, Content pages are pages that provide useful contents to the user like information about contents. Based on this consideration two types of transactions [22] are defined. The first type is auxiliary-content transactions, where each transaction including of a single content reference and all of the auxiliary references up to the content reference and another is content based. Mining on these transactions give the common traversal paths to a given content reference.

5.5 Path Completion

Client side caching gives outcomes in accessing references to those pages whose cached are not recorded in the access log. By using heuristic method, the process identifies the missing records of that session. And these are all based on site structure is called path completion [21]. For example the user return the page X which is in its current sessions, if that page is cached at client side then no request is made to server and finally no other request required. On the website we found missing reference by the knowledge of website structure and those reference information available on web server. Fig. 4 shows the missing references. The structure of a site is created by hypertext links. The structure can be obtained and pre-processed in the same way as the site of content. There should be different site structure for every server session [3].



User's actual navigation path:
P -> Q->S->T ->S -> Q -> R
What the server log shows:

URL	Referrer
P	----
Q	P
S	Q
T	S
R	Q

Fig 4: Missing Reference using path completion

6 Conclusion

Web site is considered to be the most important tool for advertisement in wide area. The feature of the website can be evaluated by examine user access of the website by web usage mining. We can identify user behaviour by the log records which is stored when user access the websites. In this paper we survey the research area of web usage mining and processing steps required for web usage mining. We have also discussed one of the processes which is data pre-processing and its various stages. Data preprocessing stages are mainly Data Cleaning, User Identification, Session Identification, Transaction Identification and Path Completion.

7 Acknowledgment

This work is supported by research grant from MPCST, Bhopal M.P., India, and Endt.No. 2427/CST/R&D/2011 dated 22/09/2011.

References

[1] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan and Mohamad Mohsin “Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm” World Academy of Science, Engineering and Technology 2008.

[2] Ms.Dipa Dixit and Ms. M. Kiruthika”Preprocessing of Web Logs” (IJCSE) International Journal on Computer Science and Engineering Volume 02, 2010.

[3] Arshi Shamsi, Rahul Nayak, Pankaj Pratap Singh and Mahesh Kumar Tiwari “Web Usage Mining by Data Preprocessing” IJCST Volume 3, Jan. - March 2012.

[4] T. Revathi, M. Mohana Rao and Ch. S. Sasanka “An Enhanced Pre-Processing Research Framework for Web Log Data” IJARCSSE Volume 2, March 2012.

[5] M. Malarvizhi and S. A. Sahaay.” Preprocessing of Educational Institution Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique”, 2012.

[6] Suneetha K.R and R. Krishnamoorthi” Data Preprocessing and Easy Access Retrieval of Data through Data Ware House” WCECS, Volume 1, October 2009.

[7] Khasawneh N. And Chan C.”Active user-based and ontology-based web log data preprocessing for web usage mining” IEEE/WIC/ACM International Conference on Web Intelligence, December 2006.

[8] Khasawneh N.,Shatnawi M.,Fraiwan M. “Converting Web Applications into Standard XML Web Services” The Tenth International Conference on Intelligent System Design and Applications, Dec 2010.

[9] R. Cooley, B. Mobasher, J. Srivastava,"Grouping web page references into transactions for mining world wide web browsing patterns", University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

[10] R. Kosala, H. Blockeel. “Web Mining Research: A Survey,” In SIGKDD Explorations, ACM press, 2000.

[11] Han, J. and M. Kamber “Data Mining: Concepts and Techniques”. A. Stephan. San Francisco, Morgan Kaufmann Publishers is an imprint of Elsevier, 2006.

[12] Raju. G. T. and Satyanarayana. P. S., “Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology”, IJCSNS International Journal of Computer Science and Network Security, Volume8, January 2008.

[13] Suneetha, K. R. and D. R. Krishnamoorthi "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, Volume 9, April 2009.

[14] Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani, "Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM", First International Conference on Networked Digital Technologies, 2009.

[15] Ramya C and Kavitha G, “An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network”, Fifth International Conference on Information Processing, Springer, 2011.

[16] Renata Ivancsy, and Sandor Juhasz, “Analysis of Web User Identification Methods”, World Academy of Science, Engineering and Technology, Volume 34, 2007.

[17] Ling Zheng, Hui Gui and Feng Li, “Optimized Preprocessing Technology for Web Log Mining”, International Conference on Computer Design and Applications, Volume1, 2010.

[18] Li Chaofeng,“ Research and Development of Data Preprocessing in Web Usage Mining”, International Journal of computer applications, 2011.

[19] Shaimaa Ezzat Salama, Mohamed I. Marie, “Web Server Logs preprocessing for Web Intrusion Detection”,Computer and Information Science, Volume 4, 2011.

[20] Liang Wei and Zhao Shu-hai,“A Hybrid Recommender System Combining Web Page Clustering with Web Usage Mining”, International Conference on Computational Intelligence and software Engineering, 2009.

[21] Yan Li, Boqin Feng and Qinjiao Mao, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, Volume 1, 2008.

[22] Jian Chen, Jian Yin, Tung, A.K.H. and Bin Liu, "Discovering Web usage patterns by mining cross-transaction association rules", International Conference on Machine Learning and Cybernetics, Volume 5, 2004.

[23] Yi Dong, Huiying Zhang and Linnan Jiao, “Research on Application of User Navigation Pattern Mining Recommendation”, Proceeding of the 6th World Congress on Intelligent Control and Automation, IEEE,China, June 21 – 23, 2006.

[24] Tasawar Hussain, Sohail Asghar and Nayyer Masood “Web Usage Mining: A Survey on Preprocessing of Web Log File” Center of Research in Data Engineering (CORDE) Department of Computer Science, 2010.

[25] Sanjay Bapu Thakare,Sangram and Z. Gawali "A Effective and Complete Preprocessing for Web Usage Mining" (IJCSE) International Journal on Computer Science and engineering, Volume 2, 2010.

[26] D.S. Rajput, R.S. Thakur and G.S. Thakur “Rule Generation from Textual Data by using Graph based Approach” International Journal of Computer Applications New York, USA, Nov. 2011.

[27] Aditi Shrivastava, Nitin Shukla “Extracting Knowledge from User Access Logs” International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012.

[28] Liu Wenyun, Bao Lingyun, “Application of Web Mining in E-Commerce Enterprises Knowledge Management”, International Conference on E-Business and E-Government, IEEE, 2010.

[29] Zhang Haiyang, “The Research of Web Mining in E-commerce”, IEEE, 2011.

[30] Vijayashri Losarwar, Dr. Madhuri Joshi, “Data Preprocessing in Web Usage Mining”, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) Singapore, July 15-16, 2012.

[31] R. Agrawal, R. Srikant, “Fast Algorithm for Mining Association Rule”, International Conference on Very Large Databases, Santiago, Chile, September 1994.

[32] R. Agrawal, R. Srikant ,”Mining sequential patterns”,11th International conference,IEEE Computer Society Press, Taiwan,1995.

[33] Jaideep Srivastav, Robert Cooley, Mukund Deshpande, Pang-Ning Tan,” Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”, ACM SIGKDD,Volume 1, Issue 2,Jan 2000.