Web Usage Mining: An Analysis

Mehak
ME (Computer Science & Engineering), University Institute of Engineering & Technology, Panjab University, Chandigarh, India
Mehakjain_1988@yahoo.co.in

Mukesh Kumar
Assistant Professor, Computer Science & Engineering Department, University Institute of Engineering &Technology, Panjab University, Chandigarh, India
Mukesh_rai9@yahoo.com

Naveen Aggarwal
Assistant Professor, Computer Science & Engineering Department, University Institute of Engineering &Technology, Panjab University, Chandigarh, India
navagg@gmail.com

***Abstract*—Web usage mining is research area in web mining. Web mining is an activity that focuses to discover new, relevant and reliable information and knowledge by examining the structure, content and usage of web. The major focus is on learning about web users and their interaction with websites. Web log files generated on web servers are used in order to extract web usage of different users. There are three types of web repositories: web server log, proxy server log, browser log. Analysing web logs for usage can not only provide important information to websites developers but also help in creating adaptive web sites.**

In this paper we discuss various sources of information for WUM, Methodology of web usage mining techniques which involves Data collection , Data pre-processing , knowledge discovery and knowledge analysis. Various applications of WUM are personalization, prefetching and caching, support to design and E-commerce. Major application of web usage mining is to predict future accesses. Thus, the result obtained after web usage mining can be used to improve the performance of prefetching and caching.

***Index terms*—Web usage mining, Methodology, pre-processing, clustering, classification, applications.**

I. INTRODUCTION:

World Wide Web is a huge repository of data. It has become one of the most important repository for storing, sharing and to distribute information. The expansion of web is very rapid which has provided a great opportunity to study user and system behaviour by exploring web access [5].

Data mining is the process that attempts to discover

Patterns in large data. Applying data mining techniques on web data to discover knowledge has been defined as WEB MINING [3]. It can be viewed as an extraction of structure from an unlabeled, semi structured dataset containing the characteristics of users or information respectively.

Data that is actually mined is varied and different approaches have been followed. Some researchers have applied mining techniques on the web logs maintained by the servers so as to discover user access and traversal path [3].

Web mining is categorized in three types:

A. *Web content mining:* It is the scanning and mining of text, pictures of a Web page to determine the relevance of the content to the search query. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

B. *Web structure mining:* Web structure mining is a tool used to identify the relationship between Web pages linked by information or direct link connection. The motive of web structure mining is generating structured summaries about information on web pages/webs.

C. *Web usage mining*: This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web page. Web server gathers this information automatically into the Access Log File.

Typical Sources of Data [1]:

1. Data generated automatically is stored in different types of log files such as server access logs, referrer logs, and client-side cookies.
2. E-commerce and product-oriented user events.
3. User profiles and user ratings
4. Meta-data, page attribute, page content, site structure.

doi:10.4304/jetwi.5.3.240-246

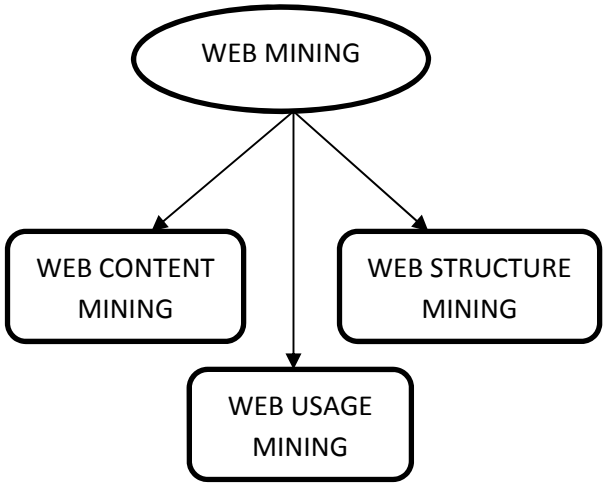

Figure 1: Types of Web Mining.

New tools promising to apply data warehousing and mining techniques on web logs have entered in the market. These include surfAid, speedTracer from IBM, bazaar analyser etc [3].

II. Web Usage Mining

Web usage mining is used to analyse web log files to discover user accessing patterns of web pages [13]. Web usage mining is a main research area in Web mining focused on learning about Web users and their interactions with Web sites. Web usage mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers. A *web log* is a listing of page reference data. A web server log file contains requests made to the web server recorded in chronological order. It is at times referred to as *clickstream* data as each entry corresponds to a mouse click [7].

Information Obtained through web usage mining [15]:

A. Number of Visitors: It is the count of users who navigates to your website and browses one or more pages on your site.

B. Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

*C. Visitor Referral Website***:** The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

D. *Number of Hits***:** This number usually signifies the number of times any resource is accessed in a Website.

*E. Time and Duration***:** This information in the server logs give the time and duration for how long the Website was accessed by a particular user.

*F. Path Analysis***:** Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.

*G. Visitor IP address***:** This information gives the Internet Protocol (I.P.) address. It is the address of the visitors who visited the website.

*H. Browser Type***:** This information provides the data of the kind of browser that was used for accessing the web site.

*I. Cookies***:** A message given to an online browser by an online server. The browser stores the message during a document known as cookie. The message is then sent back to the server whenever the browser requests a page from the server. The purpose of cookies is to spot users and probably prepare tailor-made sites for them.

*J. Platform***:** This info provides the kind of OS etc. that was accustomed access the web site.

III. Methodology of Web Usage Mining.

A web server log file contains requests made to the web server. These requests are recorded in chronological order. The popular log file formats are the Common Log Format (CLF) and extended CLF.

As shown in Figure 3[1].

Web Usage Mining includes following steps: Data Collection, Data Pre-processing, Knowledge Discovery and Pattern Analysis. As shown in Figure 4[8] and Figure 2[1].

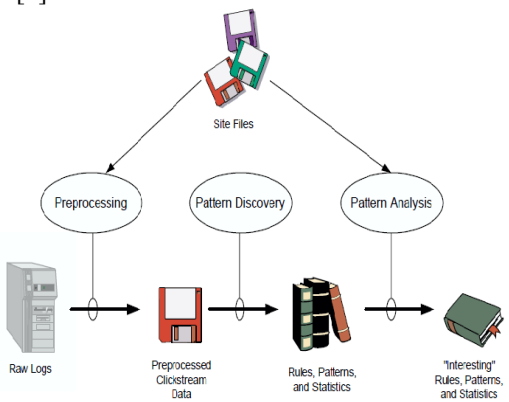

Figure 2: Basic Steps of Web Usage Mining [1]**.**

A. *Data Collection***:**

Web Usage Mining applications are based on data collected from three mainsources [13]: (i) web servers, (ii) proxy servers, and (iii) web clients [2].

i. ***Server Side:*** Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of

the request, the request line exactly as it came from the client, etc.

ii. ***Proxy Side:*** A Web proxy acts as an intermediate level of caching between client browsers and Web servers. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of group of users accessing huge groups of web servers.

iii. ***Client Side***: Most of the users have tendency to open several pages simultaneously and in between, use some non-browsing applications such as MS-word, Excel etc. for their own personal work, in such cases data recorded in server log only shows the requested time of the web pages and cannot help us to find out which web page and for how long has been really browsed on client machine. Usage data can be tracked on the client side by using JavaScript, java applets, or even modified browsers. These techniques avoid the problems of users sessions identification and the problems caused by caching (like the use of the back button). However, these approaches rely heavily on the users' cooperation and rise many issues concerning the privacy laws, which are quite strict.

*B. Data Pre-processing***:**

Some databases are insufficient, inconsistent and include noise. The pre-treatment of data is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine.

Steps involved in data pre-processing are shown with the help of block diagram below Fig 5.

i. Data Cleaning [5]:

Data cleaning is the process where irrelevant records are removed. The main aim of web usage mining is to fetch the traversal pattern; following two kinds of record are unnecessary and should be removed [5].

a. The records having filenames suffixes of GIF, JPEG, CSS and so on, which can be found incs_uri_stem field of record.

b. By examining the status field of every record in the web log, the record with status code over 299 and below 200 are removed.

ii. User and Session Identification [5]:

The main task in this step is to identify different user session from access log. A referrer-based method is used for identifying sessions. The different IP addresses distinguish different users.

<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

Figure 3: Common Log Format [1]**.**

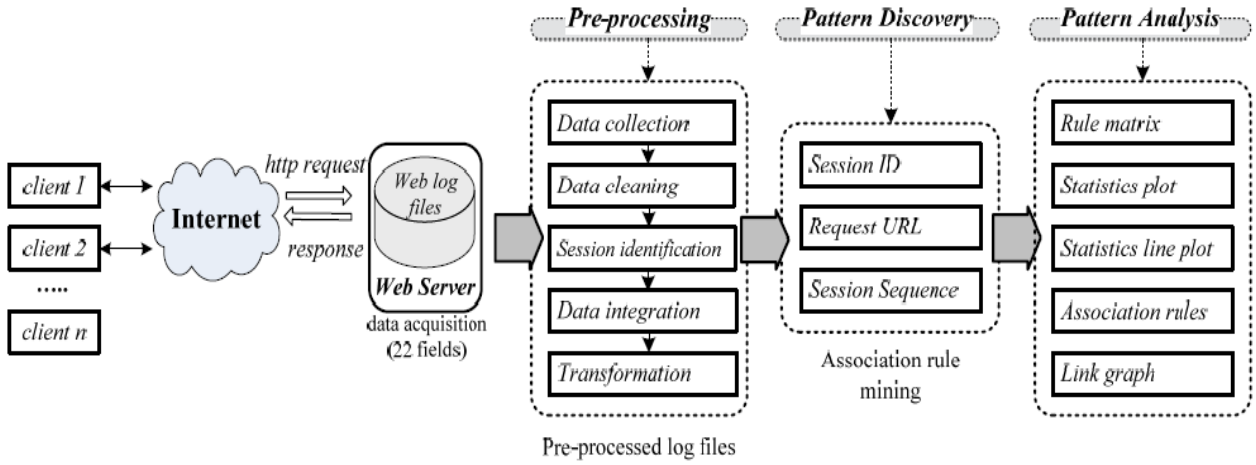

Figure 4: Algorithm Scheme for Web Usage Mining [8].

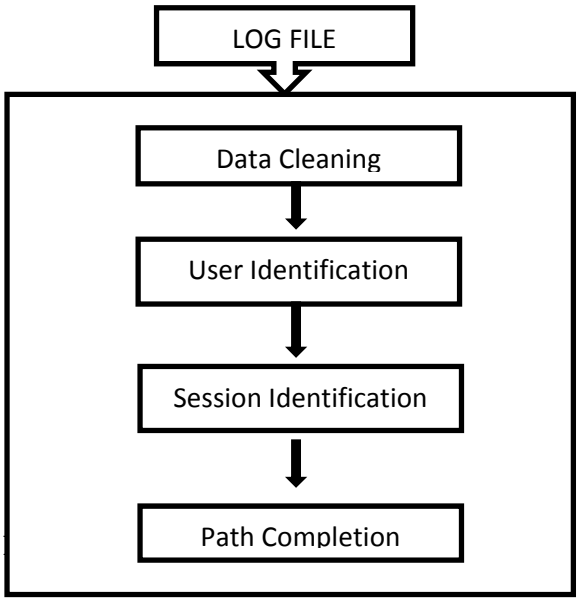

Figure 5: Steps for Data Pre-processing

a. If the IP addresses are same, then information regarding different browsers and operating systems given by client IP address and user agent indicate different users.

b. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The ReferURL (cs_referer) is checked, a new user session is identified if the URL in the ReferURL– field hasn't been accessed previously, or there is a large interval between the accessing time of this record.

iii. Path Completion:

Client- or proxy-side caching can often result in missing access references to those pages or objects that have been cached. For Example, if Page A is returned by the user during the same session, the second time when page A will be accessed again, no request is made to the server and it will result in viewing the previously downloaded version of A that was cached on the client-side. This results in the second reference to A not being recorded on the server logs. Missing references due to caching can be heuristically inferred through path completion which relies on the knowledge of site structure and referrer information from server logs [13].

Path Completion should be used acquiring the complete user access path. The incomplete access path of every user session is recognized based on user session identification. If in a start of user session, Referrer as well URL has data value, delete value of Referrer by adding '-'. Web log pre-processing helps in removal of unwanted click-streams from the log file and also reduces the size of original file by 40-50% [5].

Tools used for Pre- processing [6]:

Active Server Pages (ASP) is one of the popular scripting languages used for developing web-based application. This study focuses on this language in order to develop the application that can manipulate the server logs. To access the server logs from windows 2000, the *.dll file named *logscrpt.dll* is used to load the class object MSWC.IISLog. The MSWC.IISLog class contains several *methods* and *properties* that can be used either to retrieve log entries or write log entries [6].

In order to perform pattern mining and generalized association rules, a tool was written using Active Server Pages (ASP) to perform pre-processing techniques.

The algorithm for pre-processing is shown below [6] Figure 6:

Several attributes are ignored and the interesting fields are included in the database [6]. The algorithm that implements this function is written as [6] Figure 7:

```
Const ForReading = 1
Const ForWriting = 2
Sub ReadLog( Physical-Path, ModeFile-1,
TypeOfLogFile, ModeFile 2,StrTypeOfLogFormat)
RecordCounter = 0
Set LogReader = Server.CreateObject("IISLog")
LogReader.OpenLogFile
        LogFilePath, ModeFile-1, TypeOfLogFile,
        ModeFile-2, StrTypeOfLogFormat
LogReader.ReadLogRecord
While NOT LogReader.EndOfLogRecord
Retrieve Log Attributes
RecordCounter = RecordCounter + 1
LogReader.ReadLogRecord
Loop
LogReader.CloseLogFile
End Sub
```

Figure 6: Pre-processing Algorithm.

Approaches used for data pre-processing:

i. Pre-Processing Using Xml [5]

XML (Extended Mark-up Language) provides a structure to the records which are present in web logs. Data Pre-processing can be done using XML. Hence, understanding of web logs becomes easier. Steps involved in pre-processing using above approach are:

a. Using XML parsers DOM tree structure is created from Logs recorded in the web log.
b. Next step is user identification and session identification is same as given basic algorithm of data pre-processing.
c. Finally, the path completion helps to complete and format the paths in user session, so that these paths can be further used for analysis.
d. After the above steps, transfer the records which are present in XML file into Knowledge base.

Transfer server logs to database:

```
Declare Variables
Set DB
=Server.CreateObject("ADODB.Connection
")
Set RS
=Server.CreateObject("ADODB.Recordset"
)
ConnStr = {MsAccess Driver}
DB.OpenConnStr
RS.OpenTableName, ActiveConnection,
Add Data
RS.Update
Set Rs = Nothing
DB.Close
```

Figure7:Algorithm for transfer of log file.

ii. Pre-Processing Using Text File [5]

Data pre-processing is applied on records which are present in the web log file. Steps for pre-processing are:

a) Web log file contains log records in unprocessed form.
b) Before applying cleansing process, attributes in the text file needs to be separated using delimiter as space. These spaces help in identifying exact position of attributes/fields.
c) Steps 3 & 4 are same as in above approach.
d) After the above steps, transfer the records which are present in text file into Knowledge base.

C. Knowledge Discovery :

This is the key component of the Web usage mining. Various techniques are used to discover rules or patterns such as Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns etc.

i. Statistical Analysis :

Knowledge about visitors to a Web site is extracted with the use of Statistical techniques. Different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path can be performed by analyzing the session file. The web system report can be potentially useful for improving the system performance, enhancing the security of the System, facilitation the site modification task, and providing support for marketing decisions simply by analysing the statistical information in the report [13].

ii. Association Rules:

Association rule generation can be used to relate pages that are most often referenced together in a single server session [13].

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. These pages may not be directly connected to one another via hyperlinks. Figure 8 shows the item set generation for a set of transactions.

Most common approaches to association discovery are based on the Apriori algorithm.

This algorithm finds groups of items (page-views appearing in the pre-processed log) occurring frequently together in many transactions (i.e., satisfying a user specified minimum support threshold). Such groups of items are referred to as **frequent item sets**. Association rules which satisfy a minimum confidence threshold are then generated from the frequent item sets.

The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for prefetching documents to reduce user perceived latency.

iii. Clustering:

Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. [13].

Clustering of pages (or items) can be performed based on the usage data (i.e., starting from the user sessions or transaction data), or based on the content features associated with pages or items (keywords or product attributes).

In the case of ***content-based clustering***, the result may be collections of pages or products related to the same topic or category. In ***usage-based clustering***, items that are commonly accessed or purchased together can be automatically organized into groups.

Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users.

iv. Classification:

Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of given the class or category. Classification can be done by using supervised learning algorithms such as decision trees, naive Bayesian classifiers, *k*-nearest neighbour classifiers, and Support Vector Machines.

Classification techniques play an important role in Web analytics applications for modelling the users according to various predefined metrics.

For example, given a set of user transactions, the sum of purchases made by each user within a specified period of time can be computed. A classification model can then be built based on this enriched data in order to classify users into those

Transactions
A, B, D, E
A, B, E, C, D
A, B, E, C
B, E, B, A, C
D, A, B, E, C

Size 1		Size 2		Size 3		Size 4	
Itemset	Supp.	Itemset	Supp.	Itemset	Supp.	Itemset	Supp.
A	5	A,B	5	A,B,C	4	A,B,C,E	4
B	5	A,C	4	A,B,E	5		
C	4	A,E	5	A,C,E	4		
E	5	B,C	4	B,C,E	4		
		B,E	5				
		C,E	4				

Figure 8: Web Transaction and resulting Itemsets (minsup = 4) [16]

Who have a high propensity to buy and those who do not, taking into account features such as users' demographic attributes, as well their navigational activities.

v. *Sequential Patterns [2]***:**

Sequential Patterns are used to discover frequent sub sequences among large amount of sequential data. In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users' sessions frequently.

The typical sequential pattern has the form [14]: the 70% of users who first visited A.html and then visited B.html afterwards, in the same session, have also accessed page C.html. Sequential patterns might appear syntactically similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining.

[7] Presents a comparison of different sequential pattern algorithms applied to WUM.

D. Pattern Analysis:

The need behind pattern analysis is to filter out uninteresting rules or patterns from the set. Common form of pattern analysis consists of a knowledge query mechanism such as SQL [17]. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match certain hyperlink structure.

The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, and Usability Analysis. Visualization techniques are useful to help application domains expert analyse the discovered patterns.

IV. APPLICATIONS OF WEB USAGE MINING[2] :

The general goal of Web Usage Mining is to gather interesting information about user's navigation patterns. This information can be used later to improve the web site from the users' viewpoint. The results produced by the mining of web logs can used for various purposes [13]:

A. To personalize the delivery of web content;
B. To improve user navigation through prefetching and caching;
C. To improve web design; or in e-commerce sites
D. To improve the customer satisfaction.

A. Personalization of Web Content **:**

Web Usage Mining techniques can be used to provide personalized web user experience. For instance, in real time, it is possible to predict the user behaviour by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users.

B. Prefetching and Caching:

The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. With the use of weblogs that store user's access history can be used predict future accesses.

Typically, Web Usage Mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time.

C. *Support to the Design* :

Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. [12]. Adaptive Web sites represent a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behaviour.

D. *E-commerce* :

Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Web usage mining techniques can also be useful in Customer Relationship Management (CRM). The issues specific to business such as customer attraction, customer retention, cross sales, and customer departure are mainly in focus.

CONCLUSION:

Web Usage mining is a technique used to mine the logs available on the server. The various advantages of Web Usage Mining is to improve the structure of web page, improving the system performance and also prefetching and caching. Prefetching and pre-caching is the techniques to reduce the user's perceived latency while accessing a web page through web server. The user's access history can be used to predict its future accesses.

REFERENCES

[1] Rajni Pamnani and PramilaChawan, "Web Usage Mining: A Research Area in Web Mining", Department of computer technology, VJTI University, Mumbai, 2013.

[2] Federico Michele Facca and Pier Luca Lanzi, "Recent Developments in Web Usage Mining Research", In the proceedings of 5th international conference on Data Warehousing and Knowledge Discover, Prague, 2003.

[3] Karuna P. Joshi, Anupam Joshi, Yelena Yesha, and Raghu Krishnapuram, " Ware housing and Mining Web logs" , ACM, Pp. 63-68, 1999.

[4] Paulo Batista and Mario J. Silva, "Mining Web Access Logs of an On-line Newspaper", In the proceedings of 12th International Meeting of the Euro Working Group on Decision Support Systems, 2002.

[5] Ms.Dipa Dixit and Ms. M Kiruthika, "Preprocessing of web logs" International Journal on Computer Science and Engineering (IJCSE) Volume 02, Issue no. 07, Pp. 2447-2452, ISSN 0975-3397, 2010.

[6] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi and Mohamad Farhan Mohamad Mohsin, " Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", In proceedings of World Academy of Science, Engineering and Technology, Volume 48, December 2008.

[7] Behzad Mortazavi-Asl, "Discovering and mining user web-page traversal patterns" Master's thesis, Simon Fraser University, 2001.

[8] ResulDaş, İbrahim Türkoğlu, "Extraction of Interesting Patterns through Association Rule Mining For Improvement of Website Usability", Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006.

[9] John R. Punin , Mukkai S. Krishnamoorthy , Mohammed Javeed Zaki, LOGML: Log Markup Language for Web Usage Mining, Revised Papers from the Third International Workshop on Mining Web Log Data Across All Customers Touch Points, p.88-112, August 26, 2001.

[10] Configuration File ofW3C httpd, 1995. http://www.w3.org/Daemon/User/Config/.

[11] W3C Extended Log File Format, 1996. http://www.w3.org/TR/WD-logfile.html.

[12] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies", PHI Publication, Third edition, ISBN – 9788120330535, 2009.

[13] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD, Volume 1, Issue 2, January 2000.

[14] Eleni Stroulia Nan Niu and Mohammad El-Ramly. Understanding web usage for dynamic web-site adaptation: A case study. In Proceedings of the Fourth International Workshop on Web Site Evolution (WSE'02) IEEE, Pp. 53–64, 2002.

[15] Aniket Dash and Liju Robin George, "Web Usage Mining: An Implementation", National Institute of Technology, Rourkela, Master's report, 2010.

[16] Haizheng Zhang, Myra Spilipoulou, Bamshad Mobasher, C. lee Giles, Andrew McCallum, Olfa Nasraoui, Jaideep Srivastava, John Yen: "Advances in Web Mining and Web Usage Analysis" 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007, and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007, San Jose, CA, USA, August 12-15, 2007. Revised Papers.Lecture Notes in Computer Science 5439, Springer 2009, ISBN 978-3-642-00527-5

[17] Sathya Babu Korra, Saroj Kumar Panigrahy, and Sanjay Kumar Jena , "Web Usage Mining: An Implementation view", In the proceedings of Advances in Computing, Communication and Control International Conference, ICAC3 2011, Mumbai, India, volume 125, Pp. 131-136, 2011