Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data through Proxy Server

[1]V. Sathiyamoorthi and [2]Dr.Murali Bhaskaran

[1]Department of CSE, Sona College of Technology, Salemi-5, Tamil Nadu, India
[2]Principal,Paavai College of Engineering, Paachal, 637018, Tamil Nadu, India

Abstract
The rapid growth of the Web in terms of Web sites and their users during the last decade has put lots of pressure for Web site owners in reducing the latency of Web pages. Web caching and Web pre-fetching are two important techniques used to tackle these problems and reduce the noticeable response time perceived by users. These two techniques complement each other since the Web caching technique exploits the temporal locality, whereas Web pre-fetching technique utilizes the spatial locality of Web object. By integrating Web caching and Web pre-fetching techniques, the latency time and search space get reduced.
Due to this dramatic changes, a huge amount of data related to the user's interactions with the Web sites are recorded in the Web access log. Web access log plays an important role in predicting the user access pattern and pre-fetching and caching of Web data for better performance.
Different data mining techniques can be applied on Web usage data to mine user access patterns and this knowledge can be used in a variety of applications such as system improvement, Web site modification, business intelligence etc. This paper discusses various data preprocessing techniques that are carried out at proxy server access log which generate Web access pattern and can also be used for further applications.

Key words
Web mining, proxy server, data mining, preprocessing, pre-fetching, caching

1. INTRODUCTION

World Wide Web (WWW) has become a most powerful platform not only retrieving and sharing information, but also discovering knowledge from Web data. With the huge volume of data that are generated in day to day life, it's impossible for analyst to analyze the entire data to retrieve some useful information. To overcome this difficulty, data mining techniques are applied. Data mining is the process of retrieving knowledge from the vast amount of data stored in data repository like data base, file system and data warehouse. In order to discover knowledge from Web based application, Web mining is used. Web mining is the area in which we apply the data mining techniques to Web based applications in order to discover potential information. Web usage mining is an area of Web mining which deals with the extraction of interesting knowledge from access log produced by Web server.

Manuscript received November 5, 2011
Manuscript revised November 20, 2011

Web mining is broadly classified into three major areas that include Web content mining, Web structure mining and Web log mining or Web usage mining. Web content mining is performed on Web site content such as textual data, and the image is displayed in the Web pages. Web structure mining is performed on inter and intra Web page link in order to predict dangling link and duplicate reference etc…Web usage mining is performed on Web log file that is stored in server side. This paper discusses such data preprocessing techniques in details.
Data preprocessing is the process of removing noisy, incomplete, outlier and inconsistent data that are present in the data sets. Data mining techniques will provide incorrect results in the presence of such information. So, powerful techniques are applied to preprocess data. Data mining can be applied in various kinds of fields such as statistics, information retrieval, text mining, auto mobile and medical fields etc...
This paper is organized as follows: section 2 discusses about various related work in Web usage mining and data preprocessing activities. Section 3 gives an overview of Web usage mining, which describes data sources, techniques and applications. In section 4 data preprocessing activities like data reduction, data cleaning and related algorithms are presented and conclusion is given in section 5.

2. MATERIALS AND METHOD

More and more researchers focus on Web Usage Mining recent years (Srivastava et al., 2000; Mobasher et al., 2002; Shahabi and Kashani, 2002; Feng and Huyou, 2002). However it was Etizoni (1996) who first invented the term web mining which is concerned with extracting knowledge from web data. Web mining can be classified into three main areas such as web content mining, web structure mining and web usage mining. Web usage mining is the application of data mining techniques to large web data repositories (Cooly et al., 1997). Web structure and content mining are viewed as process based whereas web log mining is viewed as data based. Data is collected in the web server when one user accesses the web and is represented in standard formats.

The standard log format (Cooly et al., 1999; http://www.w3.org/Daemon/user/config/logging.html #common - log – file -format, consists of IP address, access date and time, request method (GET or POST), URL of page accessed, transfer protocol, success return code. In order to discover access pattern, preprocessing is necessary, because raw data coming from the web server is incomplete and only few fields are available for pattern discovery. The main objective of this study is to understand the preprocessing of usage data. On preprocessed data different techniques (http://www.w3.org/Daemon/user/config/logging.html #common - log – file -format) like statistical analysis, association rules, sequential patterns and clustering can be applied to discover user access patterns.

However, data preprocessing in Web Usage Mining has received less attention than its importance warrants. Cooly et al. (1999) presented methods for user identification, session identification, page view identification, path completion and episode identification. They proposed some heuristics to deal with the difficulties during data preprocessing. Bettina Berendt and her colleagues compared time-based and referrer-based heuristics for visit reconstruction (Joshi and Krishnapuram, 2000). They found that a heuristic's appropriateness depends on the Web site's design and on the visits' length. Doru Tanasa and Brigitte Trousse proposed advanced data preprocessing for inter sites (Fu et al., 2000). They offered the possibility of jointly analyzing multiple Web server logs. However, none of these methods are without serious drawbacks.

A number of techniques (Cooly et al., 1999) deduced from diversified fields such as statistics; machine learning, data mining, pattern recognition are applied to web usage data for pattern discovery. Statistical Analysis can be performed by a number of tools and its main goal is to give a description of the traffic on a web site e.g. most visited pages, average daily hits. Association Rules (Joshi and Krishnapuram, 2000) consider every URL requested by a user in a visit as an item and finds out the relationships between them with a minimum support level. Sequential Patterns (Fu et al., 2000) are used to discover time ordered sequence of URL's followed by past users in order to predict future ones. Clustering (http://www.w3.org/Daemon/user/config/logging.html #common – log-file -format; Srivastva et al., 2000) forms meaningful clusters of URL's by discovering similar attributes between them according to user behavior.

In some other work (Berendt et al., 2002), the authors compared time based and referrer based heuristics for visits reconstruction. Marquardt et al. (2004), Marquardt et al. presented the application of web usage mining in the e-learning area which targets on the preprocessing phase. In this context, they redefined the notion of visit from the e-learning point of view. In their approach, a learning session, visit in our case, can span over several days if this period corresponds to a given learning period.

2.WEB USAGE MINING

Web usage mining is a type of data mining process used for discovering the usage patterns from web information for the purpose of understanding and provides the requirements of web-based applications. To apply data mining techniques we need preprocessing tasks that must be performed prior to applying data mining algorithms to the data collected from server log. In this study we reviewed several data preprocessing techniques in order to identify unique users and user sessions. Also we proposed an algorithm for access matrix generation and reveal the hidden information from web log in the presence of caching based on the previous access history stored in web log file.

It was first introduced by Cooley et al (1997) and in accordance with their definition; web usage mining is the automatic discovery of user access patterns from web servers. The process of knowledge discovery and analysis of patterns focuses on user access data (web usage data). Web browsing behavior of users is captured by Web usage data from web site. In our context, the usage data is access logs on server side that keeps information about user navigation. Figure 1 shows the different phases of web usage mining.

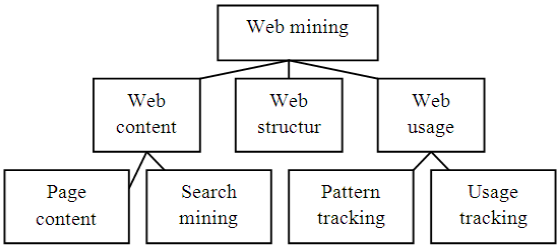

Fig. 1: Phases of web usage mining

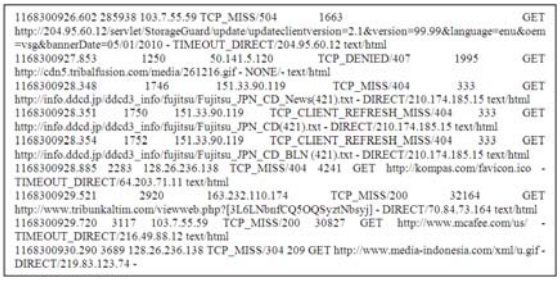

Fig. 2: Snapshot of sample proxy server log

Data source for web usage mining: Data which is used for web usage mining can be collected at three different levels (Srivastva et al., 2000).

Server level : The server stores data regarding request performed by the client. Data can be collected from multiple users on a single site.

Client level : It is the client itself which sends information to a repository regarding the user's behavior. This is done either with an ad-hoc browsing application or through client side application running standard browsers.

Proxy level : Information regarding user behavior is stored at proxy side, thus web data is collected from multiple users on several web sites, but can be used by users whose web clients pass through the proxy. In this study we will cover only web proxy server data. The information that we have at the beginning is automatically collected by web proxy server and it is stored in access log files.CERN and NCSA specified a Common Log Format (CLF) for every access stored in a log and is supported by most of the HTTP proxy servers. Figure 2 shows the example of actual log file generated by proxy server.
There are various fields in this line:

- Time stamp of the Request
- Time required to process the request in millisecond
- IP address of the machine requesting the object
- Cache Hit/Miss and Response Code
- Requested item size in bytes
- Type of method used
- Requested object name/URL
- Information if the request is redirected to another server
- Content type

The above mentioned attributes represents the minimal set of fields to be stored in every access log entry. Modern web servers like Apache and IIS permits the administrator to customize the record track of every row by inserting further variable values. User agent and referrer are the most important and if added to the CLF make up the so called combined log format (supported by Apache Web Server).

3. DATA PREPROCESSING

It is important to understand that the quality data is a key issue when we are going to mining from it. Nearly 80% of mining efforts often spend to improve the quality of data (Marquardt et al., 2004). The data which is obtained from the logs may be incomplete, noisy and inconsistent. The attributes that we can look for, in quality data includes accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility.
There is a need to preprocess data to make it have the above mentioned attributes and to make it easier for knowledge mining.The following subsections discusses about the data cleaning and data reduction algorithms.

Data cleaning : The raw proxy server log files are unsuitable for access pattern analysis. The proxy server log requires effective preprocessing to remove irrelevant data from the proxy server log file for analysis. It is important to remove all the requests from the web proxy log file that are not explicitly requested by the user. When a user requests any page using the browser, there are a number of log entries created in the log file as a page contains other web objects like, images, java script files and cascading style sheets apart from the main HTML page. The requests for these web objects are made by the browser. When we analyze the web user's behavior in the proxy server log mining; the requests to the web objects of images, java script files that the user did not explicitly request are removed. Thus, we remove all log entries that contains the following entries:

- Requests are executed by automated programs, such as web robots, spiders and crawlers.These programs generate the traffic to web sites, hence can dramatically bias the site statistics and are also not the desired category which KDWUD investigates(no spider in Proxy Server log)
- Requests for image files associated with requests for particular pages; an user's request to view a particular page often results in several log entries because that page includes other graphics, while we are only interested in what the users explicitly request, which are usually text files
- Entries with unsuccessful HTTP status codes; HTTP status codes are used to indicate the success or failure of a requested event and we only consider successful entries with codes between 200 and 299
- Entries with request methods except GET and POST

We also remove uncatchable requests from the web proxy log file, requests that contain queries. Typically query requests contain the character "?". Invalid requests from the proxy log file that refer to either Internal Server Errors with code and Server Side Errors are also removed from the proxy server log files:

//Algorithm for removing irrelevant records
//Input: Raw log file generated by proxy server
//Output: Preprocessed Log file with relevant entries consists of tuple {Timestamp, IP, URL}
//Constraint: Log file stored in Text file must transform into database for further processing

Read record in database.
For each record in database

Read fields URL Field//In proxy server Log the requested object is the URL field
If requested URL field Contains/end with Substring = {*.gif,*.jpg,*.css,*?} then
Remove records
Else if Response code is >299 or <200 then
Remove records
Else if Request method not in {GET, POST)
Remove records
Else
Save records in output
End if
Next record

Data merging: This Section present an algorithm that is used to merge different log file. At the end of the data preprocessing, the requests from all the log files recorded at different proxy server are put together into a single log file which contains the entire preprocessed request. Once we merge into a single file , it is arranged in ascending order based on timestamp of each record. This merging can be done at Internet Service Provider (ISP) to serve the need of different user communities:

//Algorithm for merging different proxy server log at ISP
//Input: Different log file generated by proxy servers
// output: Single log file containing merged output in time stamp order
For each log file in input do
Read log file one by one
Do until end of the file reached
Read record one by one from current log file.
Put it into output file
Increment the pointer to point the next record in current log file
Increment the pointer in output file to store next record
End
End
Sort the output log file entries in ascending order by the access time
Return (output)

User identification : In most cases, the log file provides only the computer address (name or IP) and the user agent (for the ECLF (http://www.w3.org/Daemon/user/config/logging.html #common - log – file -format) log files). For Web sites requiring user registration, the log file also contains the user login (as the third record in a log entry) that can be used for the user identification. When the user login is not available, each IP is considered as a user, although it is a fact that an IP address can be used by several users. For KDWUD, to get knowledge about each user's identity is not necessary. However, a mechanism to distinguish different users is still required for analyzing user access behavior. Here we are identifying frequent users and frequent pages for pre-fetching and catching of web data so that individual IP address is identified

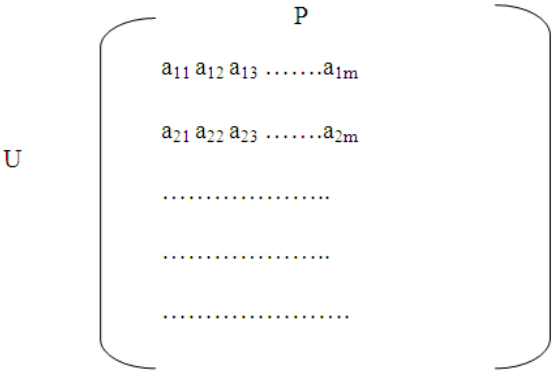

Fig. 3: Access pattern

Session identification: In Session Identification each unique IP address is identified as a different user. Session identifications process construct tuple of the form {IPi, Pagesj} i.e., set of pages visited by the particular user from the particular machine is identified. The set of URLs which is forming a session should satisfy the requirement and the time elapsed between two consecutive requests is smaller than a given Dt, which is accepted as minimum 24.5 min to 24 h (Catledg et al., 1995). The algorithm is as follows:

//Algorithm for Session Identifications
//Paramenters
//L: The set of input logs.
//|L|: The number of input logs
//Dt: Time interval//here per day i.e 24hours
//S: The set of sessions
//|S|: The number of sessions.
//Input: L,|L|, D_t
//Output: S, |S|
For each L_i of L U
If $METHOD_i$ is 'GET' and URL_i is 'WEBPAGE' //1.if
If S_k Open_Sessions with $IP_k = IP_i$ then //2.if
If $((TIME_i\text{-}END\_TIMES\ (S_k)) < D_t)$ then // 3.if
$S_k = (IP_k, PAGE_{k\ Union}\ URL_i)$
Else
CLOSE_SESSION (S_k)

OPEN_SESSION (IP_i,URL_i)
End if //end of 3.if
Else
OPEN_SESSION (IP_i,URL_i)
End if //end of 2.if
End if //end of 1.if
End for

Feature extraction: The most frequent visitors are identified and stored in vector - users $\{u_1, u_2, u_3,..,u_n\}$. The most frequent pages visited are identified and stored in a vector - pages $\{p_1, p2 \ldots p_m\}$. For each user u_i the number of visits made to each page p_i is then collected and stored in a vector called access pattern that is shown in Figure 3. Thus, entries in a_{ij} in the access pattern vector indicates the number of times user i has visited the page j. Thus, we can say that this vector gives information about the preferences of each user visiting our site:

//log_database contains the preprocessed log file entry
//Query for Identifying Frequently Visited Pages

Select page, count (user) as No_of_Visits from log_database group by page having count (user)>2:

//Query for Identifying Frequently Visited User

Select user, count (page) as No_of_Visits from log_database group by user having count (page)>2:

//Query for Identifying User, Pages and no of times each user is accessing the pages

Select user, page, count (page) as No_of_Visits from log_database group by user, page having count (page)>2.

U-User and P-Page, a_{11} is the no. of times user u_i have been accessing the page p_i:

//Algorithm for binary Access Matrix Generation
//Input: Integer Matrix A and threshold T
//Output: Boolean Access Matrix A
Step 1: Calculate sum of each row that is i.e., Tot=$\sum$A1j
Step 2: Divide each row element of the matrix by corresponding row sum that is A_{ij}/Tot.
Step 3: If each entry in the matrix is greater than the threshold value T then replaces it by '1' else replaces it by '0'

From the above matrix 1 represents the page is most frequently accessed by particular user and 0 represents the not frequently accessed:

$$a_{ij} = \begin{cases} 1 & a_{ij} > x \\ 0 & \text{otherwise} \end{cases}$$

where, a_{ij} - no. of visits made by i^{th} user to j^{th} page. Hence, each row contains the access pattern of the users visiting the site. Each row is then normalized such that:

$$\Sigma_j a_{ij}=1/0$$

The normalization eliminates the effect of the magnitude of the vector during comparisons. The normalized vector is then given as input to the clustering module to group the user access patterns based on their similarities. Since, ART1 network is used for clustering purpose; the normalized user access patterns are binary encoded is shown in fig 3.

4. APPLICATIONS

Personalization: Personalization is achieved by keeping track of previously accessed pages for E-Commerce (Pirolli et al., 1996). Making dynamic recommendations to a web server on the basis of user profile in addition to usage behavior is very attractive to many applications (Buchner and Mulvenna, 1998). Web usage mining is an excellent approach for achieving this objective as described in (Grossman and Frieder, 2004) existing recommendation systems.

System improvement: Response time and performance are crucial to user satisfaction from services like Web base applications, databases and networks. Similar qualities can be expected from the user of web services. Web log mining could provide the key to understand web traffic behavior, which can in term be used for developing policies for web coaching, Network Transmission (Anderson, 2002).

Site modification: The pleasant appearance of a web site, in term of both content and structure, are important factors to many applications like product catalog for E-Commerce. Web usage mining provides detailed feedback on user behavior and it can provide the web site designer information. This information can be used to take redesign decisions. In adaptive website (Diebold and Kaufmann, 2001; Anderson, 2002) structure of a Website changes automatically on the basis of usage patterns discovered from server logs.

Business intelligence: Information on how customers are using a website is crucial for marketers of e-tailing businesses. Buchner and Mulvenna (1998) and Tan and Kumar (2000) has presented a knowledge discovery process to discover marketing intelligence from web data. Four distinct steps in customer relationship life cycle are identified which can be supported by their knowledge discovery techniques: Customer attention, Customer retention, cross sales and customer departure.

5. RESULTS AND DISCUSSION

The data sets for testing have been obtained from IRCACHE. IRCache is a NLANR (National Laboratory of Applied Network Research) project that encourages web caching and provides data for researchers. The files that have been used for the testing of the scheme can be obtained under: ftp.ircache.net/Traces/ bo2[1].sanitized-access.2007-01-09/. The proxy server log files used for testing are bo2 [1].sanitized-access.2007-01-09 and bo2[1].sanitized-access.2007-01-09. These represent the traces for proxy server installation at Research Triangle Park, North Carolina for the dates 01/09/2007 and 01/10/2007. In Table 1, each of the dataset contains access patterns for about 117 and 118 different client IP addresses. Figure 4 shows the number of time each user accessing the corresponding URL. In Table 2 we have compared proposed work with the existing work done by the different researchers.

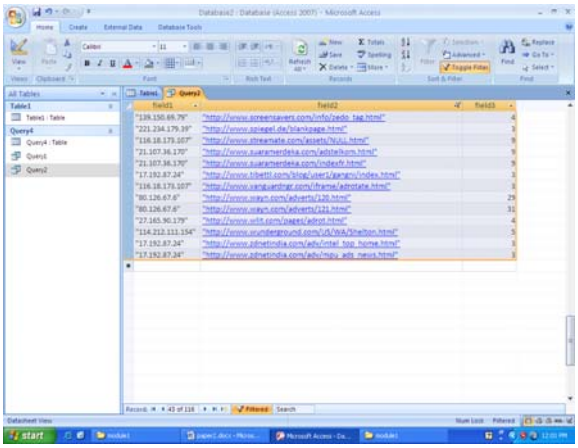

Fig. 4: Access pattern generation

Table 1: Results of data preprocessing

Data source name	Duration	Size of the data source	Size of Data source after preprocessing	No. of unique user	No of unique pages	No of frequent user	No of frequent page
bo2[1].sanitized-access.20070110	24 hours 1/9/2007	30.2MB	1.41MB	63,	2545	48	122
bo2[1].sanitized-access.20070109	24 Hours 1/10/2007	35.4MB	1.43MB	63,	2545	54	280

Table 2: Comparative study of data preprocessing with proposed work

	Data source			Data cleaning				Data formatting and structuring		
Related work	Site map	Site semantic	Multiple servers	Merging	Anonym -ization	Removing images	Removing web robots	User ID	Session ID	Generalization and Aggregation
Berendt *et al*. (2002)						Y			IP	
Chen *et al*. (1998)	Y									
Srivastva *et al*. (2000)	Y				Y	Y	Y	Login	IP, Agent	
Fu *et al*. (2000)						Y			IP	Y
Joshi and Krishnapuram (2000)						Y			IP	
Shahabi and Kashani (2002)	Y				Y				IP, Session ID	
Marquardt *et al*. (2004)						Y		Login	Login	
Proposed Work	Y		Y	Y		Y	(Not Required) Because we are using Proxy Log)	IP	IP,Page	Y

DISCUSSION

In this world, the necessity of fast information access is well known. Wide variety of information is available on line. By applying mining techniques in proposed system, online information providers can predict the user needs and provide information at a much faster rate. Thus, they can satisfy and retain their visitors. This is just one of the applications where the proposed system of the project can be used. This system can also be used with small modifications for the following applications:

- Predicting user purchase pattern of commodities
- Predicting sales pattern of commodities
- Predicting the various other links that a user might be interested
- Modifying the design of a site according to user taste

REFERENCES

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. Web Usage Mining:Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000, Vol.1 (2):1-12

[2] Mobasher B., Dai H., Luo T, Nakagawa M. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery, 2002,6(1): 61-82
[3] Shahabi C., Kashani F.B. A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking. Proc. WEBKDD 2001: Mining Web Log Data across All Customer Touch Points, LNCS 2356, Springer-Verlag, 2002: 113-144.
[4] Zhang Feng, Chang Huyou. Research and development in Web usage mining system-key issues and proposed solutions: a survey. Machine Learning and Cybernetics, 2002(2):986-990.
[5] O.Etizoni. The World Wide Web: Quagmire or Gold Mine. Communications of the ACM, 39 CII): 1996, 65-68.
[6] Robert Cooly, Bamshad Mobasher, Jaideep Srivastava (1997) : Web Mining Information and Pattern Discovery on the World Wide Web,1-10
[7] Robert Cooly, Bamshad Mobasher, Jaideep Srivastava (1999): Data Preparation for Mining World Wide Web browsing Pattern.5-32.
[8] http://www.w3.org/Daemon/user/config/logging.html #common - log – file -format.
[9] Jaideep Srivastva, Robert Cooly, Mukand Deepande, Pang-MingTan (2000): Web Usage Mining: Discovery and Applications of usage Patterns from Web Data.Vol.1 issues 2, 12-20
[10] A. Joshi and R. Krishnapuram. On Mining Web Access Logs. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 2000, 63- 69
[11] Y. Fu, K. Sandhu, and M. Shih. A Generalization-Based Approach to Clustering of Web Usage Sessions. In Proceedings of the 1999 KDD Workshop on Web Mining,San Diego, CA, vol. 1836 of LNAI,. Springer, 2000, 21-38
[12] M. S. Chen, J. S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. Knowledge and Data Engineering, 10(2), 1998, 209-221
[13] Peter Pirolli, James Pitkow and Ramna Rao.Silk from and Sow's Ear : Extracting usable Structure from the Web. In CHI – 96 Vancouver, 1996,118-125.
[14] Alex Buchner and Maurice D Mulvenna. Discovering Internet Marketing Intelligence through online Analytical Web Usage Mining. SIGMOD record, 27(4) : 1998,54-61.
[15] David A. Grossman and Ophir Frieder. Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series) (2nd Edition) (Paperback - Dec 20, 2004).
[16] Boris Diebold and Michael Kaufmann. usage based Visualization of Web Localities. In Australian Symposium on Information Visualization Pages ,2001,159-164.
[17] Corin R. Anderson: A Machine Learning Approach to Web Personalization Ph. D. Thesis, University of Washington,2002.
[18] Pang-Ning Tan and Vipin Kumar. Modeling of Web Robot Navigational Patterns. In WEBKDD 2000 – Web Marketing for E-Commerce-challenges and Opportunities, Second International Workshop August 2000.
[19] B. Berendt, B. Mobasher, M. Nakagawa and M.Spiliopoulou. The Impact of Site Structure and User Environment and Session Reconstruction in Web usage analysis. In proceedings of the forth Web KDD 2002 workshop at the ACM–SIGKDD Conference on Knowledge Discovery in Databases (KDD 2002), Edmonton, Alberta, Canada, 2002.159-179
[20] C. Marquardt, K. Becker, and D. Ruiz. A Preprocessing Tool for Web usage mining in the Distance Education Domain. In Proceedings of the International Database Engineering and Application Symposium (IDEAS' 04), 2004, 78-87.
[21] Catledg, L; Pitkow, J.: "Characterizing Browsing Behaviors on the World Wide Web", In Computer Networks and ISDN System 27(E) 1995, 1065-1073