Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Pre-Processing E-Commerce Web Log Files for Web usage Mining

Preeti Gupta
Computer science Engineering,
Amity School of Engineering, Noida
India

Abstract— Recently, the web is becoming an important part of people's life. It has become best platform to run successful businesses. Therefore, Selling products or services online have become seamless and hassle free. Web usage mining, data mining techniques is applied to discover patterns from the Web usage data in order to understand and better serve the needs of Web-based applications. There are several preprocessing tasks that must be performed prior on data collected from server log for data mining algorithms to apply. Data preprocessing for data mining are algorithms to implement on the data necessary to further the process of transforming raw data into abstraction. This paper explains several data preparation techniques that pre-process web log files in order to identify unique users and user data session which can be used to improve the performance features of web mining.

Keywords— pre-processing, web usage mining, user identification, session identification, data cleaning

I. INTRODUCTION

In last 10 years E-commerce and has changed the face of competition in most of the enterprises. Advancement in internet technologies has seamlessly automated interface processes among customers, retailers, distributors, and manufacturers. In general E-commerce have enabled on-line transactions and shopping a seamless activity. E-commerce applications are producing volume of data daily and since hundreds of new users are daily joining one or another e-commerce website or application, it foremost concern to understand and better serve the needs of them.

Web usage mining is an application that uses data mining to analyze and discover interesting patterns of user's usage of the web. This records the user's behavior when the user browses or makes transactions on the web site. Analyzing data through web usage mining can help effective Web site management, creating adaptive Web sites, business and support services, personalization, and network traffic flow analysis and so on.

Web usage mining includes three phases namely pre-processing, pattern discovery and pattern analysis. Data pre-processing is to change a web data into reliable data. Pattern discovery is used to find interesting patterns using techniques like Path Analysis, Association rules, Clustering, Classification etc. Pattern analysis phase uses techniques such as OLAP/ Visualization Tool, Knowledge Query Management, Intelligent Agents For multidimensional analysis & Decision making.

Lots of research has been done in pattern discovery and analysis steps but less emphasis is paid on data preprocessing step. Pre-processing of data an important step in the web usage mining as this step makes data suitable for web mining. The phrase "garbage in, garbage out" is well suited to data mining process. In most of the E-commerce applications, data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g. salary: -20,000), invalid data sets (e.g. Age: 8, Have-Child: Yes), missing values (e.g. region: N/A), etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis[1].

Efforts to convert raw data into abstraction have been explained in this work. First section gives brief of web usage mining, pre-processing and its steps, various types of web log files, their sources and format. Second section explains various methods and algorithms to remove inconsistent data, filling missing values, identifying session and identifying user as pre-processing steps. Third section give brief of problems faced in this work, and last section draws conclusions and future recommendations.

II. WEB USAGE MINING

Web usage mining also known as web log mining, process to discover interesting and frequent user access patterns from web browsing data that are stored in web server logs, proxy server logs or browser logs. Web usage mining is the application that uses data mining to analyse and discover interesting patterns out of user's usage data on the web. This records the user's behaviour when the user browses or makes any transactions on the web site[2]. It is an activity that involves the automatic discovery of patterns from one or more web servers log files.

III. WEB DATA

User's usage data can be collected from the server- side, client-side, proxy servers if used, or from company's database. Each type of data collection differs not only in terms of the location of the data source, but also the type of data available,

the section of population from which the data was collected, and method of implementation. The usage data collected at the different sources will represent the navigation patterns of different sections of the overall web traffic, ranging from single-user, single-site usage to multi-user, multi-site access.

a) Server Level Collection: A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. These log files can be stored in various formats such as Common log or Extended log formats.

However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log. In addition, any important information passed through the POST method will not be available in a server log. Packet sniffing technology is an alternative method to collecting usage data through server logs. Packet sniffers monitor network traffic coming to a Web server and extract usage data directly from TCP/IP packets. The Web server can also store other kinds of usage information such as cookies and query data in separate logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. The Web server also relies on other utilities such as CGI scripts to handle data sent back from client browsers.

b) Client Level Collection: Client-side data collection can be implemented by using a re-mote agent (such as JavaScript or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities. The implementation of client-side data collection methods requires user cooperation, either in enabling the functionality of the Javascripts and Java applets, or to voluntarily use the modified browser. Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems. However, Java applets perform no better than server logs in terms of determining the actual view time of a page. In fact, it may incur some additional overhead especially when the Java applet is loaded for the first time. JavaScript, on the other hand, consumes little interpretation time but cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user, single-site browsing behaviour. A modified browser is much more versatile and will allow data collection about a single user over multiple Websites. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities. This can be done by offering incentives to users who are willing to use the browser, similar to the incentive programs offered by companies such as NetZero and All Advantage that reward users for clicking on banner advertisements while surfing the Web.

c) Proxy Level Collection: A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterising the browsing behaviour of a group of anonymous users sharing a common proxy server.

IV. PHASES OF WEB USAGE MINING

The process of web usage mining consists of three important teps: (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis as shown in figure 2.2.

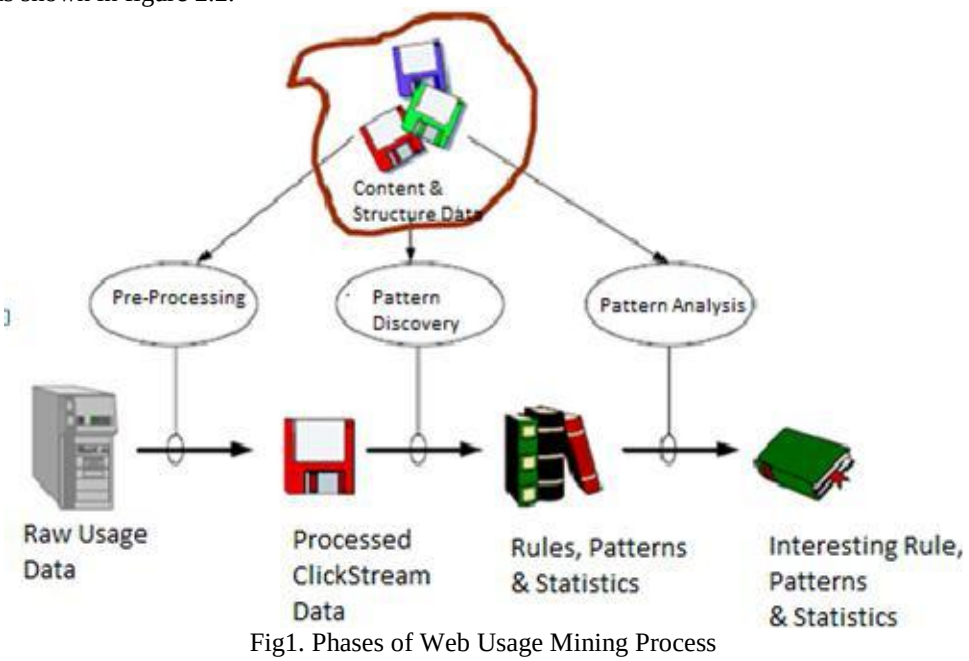

Fig1. Phases of Web Usage Mining Process

a) Pre-Processing: The purpose of Data Pre-processing is to change a web data mining into reliable data. If there is much irrelevant and redundant information present or noisy and unreliable data, then web usage mining result can be irrelevant and different from expected output. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The output of data this phase is the final training set for web mining[3]. The outputs are the user session file, transaction file, site topology, and page classifications. It's always necessary to adopt a data cleaning techniques to eliminate the impact of the irrelevant items to the analysis result. The usage preprocessing probably is the most difficult task in the Web Usage Mining processing due to the incompleteness of the available data. [5]

b) Pattern discovery: Pattern discovery converges the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. Various techniques and method used for pattern discovery are statistical analysis, classification, association rules, clustering, sequential pattern, dependency modelling.

c) Pattern Analysis: Pattern Analysis is a last phase of the Web usage mining process. The goal of this phase is to eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of Web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. This can be done with the help of some analysis methodologies and tools. There are two most common approaches for the pattern analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations [4]. All these methods assume the output of the previous phase has been structured. There are more techniques coming out in recent years, such as visualization etc.

All these phases of web usage mining have their importance in complete process. A lot of work and research[7] have been done and going on last two phases of process but less attention is given to pre-processing phase though it is most important as result always depends on how complete, accurate and reliable input was. Thus let us discuss pre-processing its methods and some approaches of data pre-processing.

V. PRE-PROCESSING

Data Pre-processing is the most important step of the entire data mining process. The goal of the data preprocessing is to turn the web logs into some reliable, complete and accurate sources to satisfy the need of web mining algorithms. Statistics show that in data mining processes the process of preprocessing accounts for 60% of the entire workload. Because the data from real world are often incomplete and inconsistent with the noise, the data preprocessing can improve the quality of data for web usage mining. It not only save a lot of time and space for next mining tasks but also plays an important role for decision making and forecasting[6]. The typical problem is distinguishing among unique users, user sessions, transactions etc[17].

Pre-Processing Steps: Data pre-processing includes 5 steps:

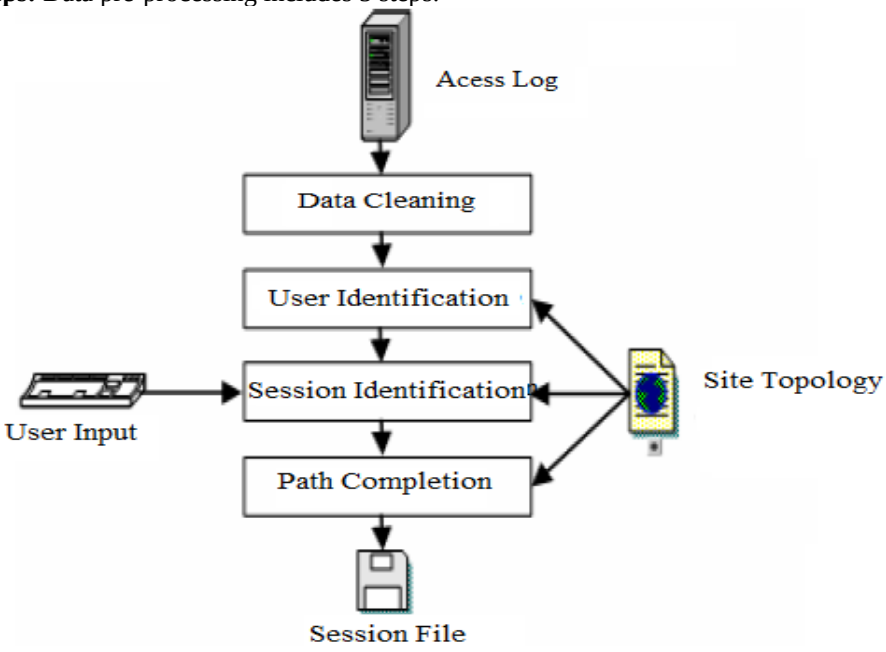

Fig2. Pre-Processing Steps

VI. DATA CLEANING

The first steps of data preprocessing is to remove useless, irrelevant and redundant log entries. Typically, the process concerning non-images, multimedia files, page style files, JavaScript files, and removing web robots' requests[8].

There are three kinds of irrelevant or redundant data to be removed. They are:

- **Additional Requests:** Graphics and other scripts files are requested in addition to the HTML file, because of the stateless behaviour of the HTTP protocol. Since aim of Web Usage Mining is to get a visualization of the user's web usage, it does not make sense to consider file requests that the user did not requests explicitly. Eliminate entries with suffixes like gif, jpeg, gif, jpg, css etc.

- **Entries with error:** Status code shows the success or failure of a request. Entries with status code less than 200 and greater than 299 are failure entries which can be removed.

- **Removing Web Robots' Requests:** Web robot, also called spider or bot is a tool that hit the website regularly to access its contents. WRS automatically follow all links of a website. Search engines like Google, WRS regularly used to collect all pages from website to update their search indexes. Hit of queries from one WR may be equal to the number of URI's website. If the Web site does not attract many visitors, the number of inquiries come from all WRS who have visited the site could be more than human-generated requests. Therefore, a WR will generate a huge number of requests on a Web site. Moreover, the requests of a WR are out of the analysis scope, as the analyst is interested in discovering knowledge about users' behavior. Most of the web bots identify themselves with the user agent field of the log file. Several reference databases known robot is maintained. However, these databases are not exhaustive and every day new WRS show or a new name, making the WR identify task more difficult. To identify WR hosts, we currently use three heuristics:

  a) **Robots.txt (RT):** We look for all the hosts that have requested the page /robots.txt. This file contains browsing rules for the WRs that index the Web site, such as the names of the folders not to be indexed.

  b) **Known User Agent (UA):** We use a list of user agents known as robots. The list is created using data from various sources such as[10,11].

  c) **Robot IP (RI):** We use a list of Robot IP know as robots. The list is created using data from various sources such as[8,13]

Once all e Web robots entries are identified, we can remove them. This procedure is straightforward, consisting in the removal of all the requests issued by pairs of (Host, User Agent) identified as being a Web robot. Removing WR-generated log entries simplifies the task that will follow; also it removes irrelevant sessions from the log file.

By filtering out useless data, we get log files requiring reduced storage space. For example, by filtering out the image entries itself, we at 50% of its original size to cut down the web server log files size. In data cleansing step required attributes like date, time, IP address, User Agent [9], URL requested, URL referred, time taken are selected for next steps reduce the processing time. So Attribute Subset Selection is done. Following shows algorithm for data cleansing step in pre-processing phase of web usage mining process.

Algorithm Name: Data Cleaning of Web Log File

Input: Web Server Log file
Output: Cleaned web log file

Step1: Read log entry from log file
Step2: If (entry has resource request with extension (.gif, .jpeg, .jpg, .css) OR (different error like HTTP 404 or more) found
then remove this entry from file
End of If.
Step3: Repeat the steps 1 and 2 until EOF encountered
Step4: End the process.

VII. USER IDENTIFICATION

A user is identified as the principal who uses a client to retrieve and provide the resources. In this work, we used the following heuristics to identify the user:

- Each IP address represents one user

- If multiple entries have same IP address, but the agent log shows a change in User Agent (with Version) or operating system (with screen resolution and processor speed etc.), an IP address represents a different user.

So user identification is done by IP address, agent, cookies or user registration. Following is algorithm used for identifying number of distinct user from cleaned web log file.

Algorithm Name: User Identification

Input: Processed Web Log File.

Output: Number of unique user.

Step1: Read entry from web log file.
Step2: User's IP addresses of two consecutive entries are compared.
Step3: If (IP address is same) then check user's browser and operating system
if both are same then
consider same user
else
consider new user
end if
end if
Step 4: Repeat above 2 steps until EOF encountered.

VIII. USER SESSION IDENTIFICATION

A user session identified one or more sessions over the web servers. The target user clicks (click stream) means a delimited set of individual sessions each user accesses the page divide. The methods to identify user session include time out mechanism.

The following is the rules we use to identify user session in our experiment:

- If there is a new user, there is a new session
- If the time between page requests exceeds a certain limit (30 or 25.5mintes), it is assumed that the user is starting a new session

User Session is also used to find out all resource references by user of application. We differentiate the entries into different user sessions through a session timeout. If the time between page requests exceeds a certain limit, it is assumed that other user-session has started. We have used 5 minute timeout for session's timeout property value. Following algorithm is used for identifying number of session from web log file.

Algorithm Name: Session Identification from web log file

Input: Web Server Log file
Output: Number of Session

Step1: sessions = {};
Step2: users = {};
Step3: N = 0
Step4: While not EOF(web_log_file) DO
Step5: logEntry = Read(web_log_file)
Step6: If(logEntry.TimeTaken > 5) OR (logEntry.UserId not in users) then
Step7: N = N+1;
SK = logEntry.URL
sessions = sessions U {SK}
Write(SessionFile , sessions)
End If
End While
Thus after completion these Preprocessing steps, we have cleaned Web Server Log file.

IX. PATH COMPLETION

In order to reliably identify unique user session, it should determine if there are important accesses that are not recorded in the access log. This is next step in data preprocessing called path completion. There is some reason for path incompletion, for example, local cache, the cache agent, "post" technique and the browser's "back" button to access some important log file accesses can be result of log entries not done, and There are number of Uniform Resource Locators (URLs) entered in the log can be less than the real one. Use of local caching and proxy server path to meet the production difficulties because users use the server logs record without leaving any local caching or proxy server caching the pages can use[14] and mechanisms such as local caches and proxy servers can severely distort the overall picture of user traversals through a Web site.

Current methods to try to overcome this problem include the use of cookies, cache busting, and explicit user registration. As detailed in [18], none of these methods are without serious drawbacks. Cookies can be deleted by the user, cache busting defeats the speed advantage that caching was created to provide and can be disabled, and user registration is voluntary and users often provide false information[17].

Methods similar to those used for user identification can be user for path completion. To accomplish this task needs to refer to referrer log and site topology, along with temporal information to infer missing references. Of the referred URL of a requesting page does not exactly match the last direct page requested, it means that the requested path is not complete. Furthermore, if the referred page URL is in the user's recent request history, we can assume that the user has

clicked the "backward" button to visit page. But if the referred page is not in the history, it means that a new user session begins, just as we have stated above. We can mend the incomplete path using heuristics provided by referrer and site topology.

X. TRANSACTION IDENTIFICATION

Before any mining is done on web usage data, sequences of page references must be grouped into logical units representing web transactions. A transaction differs from a user session in that the size of a transaction can range from a single page reference to total number of page references in a user session, depending on the criteria used to identify transactions. Unlike traditional domains for data mining, such as point of sale databases, there is no convenient method of clustering page references into transactions smaller than an entire user session. This problem has been addressed in[19, 21].

How to define transaction depends on what kind of knowledge we want to mine. In[21, 22] , user session mentioned above as user session based on duration and transaction here as user session based on structure. As Web Usage Mining aims at analyzing navigation patterns of users, so it is reasonable to take session as transaction. However, WebMiner[21,19], concentrates on association rule and sequential pattern mining, so it tends to divide user session into transaction, which is set of semantically related pages.

In order to divide user session into transaction, WebMiner classifies pages in a session into auxiliary pages and content pages. Auxiliary pages are those that are just to facilitate the browsing of a user while searching for information. Content pages are those that user are of interest and that they really want to reach. Using the concept of auxiliary and content page references, there are two ways to define transactions. The first would be to define a transaction as all the auxiliary references up to and including each content reference for a given user, which is a so-called auxiliary-content transaction as all of the transaction. The second method would be to define a transaction as all of the content references for a given user, called content-only transactions. Based on the two definitions, WebMiner employs two methods to identify transaction: one is reference length; the other is maximal forward reference.

In this work, to implement transaction identification, the user sessions and the user access paths are extracted from the Web access log and missing information is appended. These tasks are accomplished with the application of the referrer-based method, which is an effective solution to the problems introduced by using proxy servers, local caching and firewall. Meanwhile, the reference length of accessed pages is calculated with the consideration of the time spent on data transfer over Internet. Then two kinds of transactions are defined, i.e. travel-path transactions and content-only transactions. These two kinds of transactions are constructed by the maximal forward references (MFR) algorithm and the reference length (RL) algorithm, respectively. As verified by practical Web access log, it is shown that the transactions can be efficiently identified while the reliability of the original Web access data is obviously improved for the further researches[15].

XI. CONCLUSION

This work that the use various preprocessing steps to clean, and complete the web log file for web usage mining, display the details of preprocessing steps applied to the data. Preprocessing not only reduces log file size, but also increases the quality of data available. It provides quality, reliable and concise but complete data for more accurate and fast web mining process.

However, many problems remain such as data collection, applications of some heuristics in some phases of data pre-processing, the accuracy of user identification and session identification, applying the results of data pre-processing to patterns discovery and so on. We'll focus on solving these issues in the future.

REFERENCES

[1] Pyle D., 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.

[2] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, June 2013. Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery. International Journal of Data Mining Techniques and Applications, ISSN: 2278-2419.

[3] Ankit R Kharwar, Chandni A Naik, Niyanta K Desai, October 2013. A Complete PreProcessing Methodfor Web Usage Mining. International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459 Volume 3, Issue 10.

[4] O. Zaiane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by applying OLAP and Data Mining Technology on Web Logs. In Advances in Digital Libraries, pages 19-29, Santa Barbara, CA, 1998

[5] Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000 R. Cooley. WebSIFT: The Web Site Information Filter System

[6] Zhu, H.: Data Preprocessing Algorithm of Web Log Mining. China Master's Theses Full-text Database (August 2010)

[7] Han, J., Meng, X., Wang, J.: The Research of Web Mining. The Research and Development of Computer 38(4), 405–414 (2001)

[8] Robots IP Address", http://chceme.info/ips/

[9] Andreas Staeding, "User-Agents (Spiders, Robots, Crawler, Browser)", http://www.user-agents.org/

[10] Andrew Shen , "Http User Agent List", http://www.httpuseragent.org/list/

[12] Andreas Staeding , "User-Agents (Spiders, Robots, Crawler, Browser)", http://www.user-agents.org/
[13] Volatile Graphix, Inc.",http://www.iplists.com/nw/
[14] Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, IEEE,2008.
[15] Yan Li, Bo-qin Feng, June 2009. The Construction of Transactions for Web Usage Mining. Computational Intelligence and Natural Computing, 2009. CINC '09. International Conference on (Volume:1).
[16] Zhang Huiying, Liang Wei. An Intelligent Algorithm of Data Pre-processing in Web Usage Mining, Proceedings of the 5th World Congress on Intelligent Control and Automation, 2004.
[17] Ke Yiping , Dec 2003.A survey on Preprocessing Techniques in Web Usage Mining , The Hong Kong University of Science and Technology .
[18] Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: Extracting usable structures from the web. In Proceedings of 1996 Conference on Human Factors in Computing -11 Systems (CHI-96), Vancouver, British Columbia, Canada, 1996.
[19] M.S. Chen, J.S. Park, and P.S. Yu, 1996. Data mining for path traversal patterns in a web environment. In Proceedings of the 16th International Conference on Distributed Computing Systems, pages 385-392.
[20] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, November 1997. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns (1997), in Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX-97).
[21] B. Berendt, M. Spiliopoulou. Analyzing navigation behavior in Web sites integrating multiple information systems. VLDBJournal, Special Issue on Databases and the Web 9, 1 (2000), 56-75.

