International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 279-283

AN ALGORITHMIC APPROACH TO DATA PREPROCESSING IN WEB USAGE MINING

Navin Kumar Tyagi[1], A.K. Solanki[2] & Sanjay Tyagi[3]

Web usage Mining is an area of web mining which deals with the extraction of interesting knowledge from logging information produced by web server. Different data mining techniques can be applied on web usage data to extract user access patterns and this knowledge can be used in variety of applications such as system improvement, web site modification, business intelligence etc. Web usage mining requires data abstraction for pattern discovery. This data abstraction is achieved through data preprocessing. In this paper we survey about the data preprocessing activities like data cleaning, data reduction and related algorithms.

Keywords: Web Usage Mining, Web Server, Data Mining, Data Preprocessing

1. Introduction

Web has recently become a powerful platform for, not only retrieving information, but also discovering knowledge from web data. Historically, the conception of discovering useful patterns in data has been given a variety of names like data mining, knowledge extraction, Information discovery, Information Harvesting, data Archeology, and data pattern processing. It was Etzioni [1] who first invented the term web mining which is concerned with extracting knowledge from web data. There has been huge interest of Researchers towards web mining. On the basis of definition of web mining two different approaches can be proposed. One is process based and other is data based. Data based application is more widely accepted today. In this prospect, web mining is the application of data mining techniques to extract knowledge from web data, where structure (hyperlink) or content (actual data in web pages) or usage data (web log data) is used in the mining process.

On the basis of web data three categories of web mining are proposed, which are web structure mining, web content mining & web usage mining. Web usage mining is the application of data mining techniques to large web data repositories [2]. Data is collected in web server when user accesses the web and might be represented in standard formats. The log format of the file is CERN (Common log formats)[3], which consists attributes like IP address, access date and time, request method (GET or POST), URL of page accessed, transfer protocol, success return code etc. 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. Main objective of this paper is to understand the preprocessing of usage data. On preprocessed data different techniques [4] like statistical analysis, association rules, sequential patterns and clustering can be applied to discover user access patterns.

This paper is organized as follows. In section 2 an overview of web usage mining is given which describes data sources, techniques and applications. In section 3 data preprocessing activities like data reduction, data cleaning and related algorithms are presented. In section 4 we discuss some related work and conclusion is given in section 5.

2. Web Usage Mining

The term web usage mining was introduced by Cooley et al. in 1997 and in accordance with their definition; web usage mining is the automatic discovery of user access patterns from web servers. The process of 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.

2.1. Data Source for Web Usage Mining

Data which is used for web usage mining can be collected at three different levels [5].

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

[1]Department of Computer Science Engg. & I.T., M.I.T., Bulandshahr, INDIA

[2]Director, Meerut Institute of Engineering and Technology, Meerut, INDIA

[3]Department of Computer Science & Applications, K.U.,Kurukshetra, INDIA

Email: [1]nt_1974@rediffmail.com, [2]solankimiet09@gmail.com, [3]tyagikuk@gmail.com

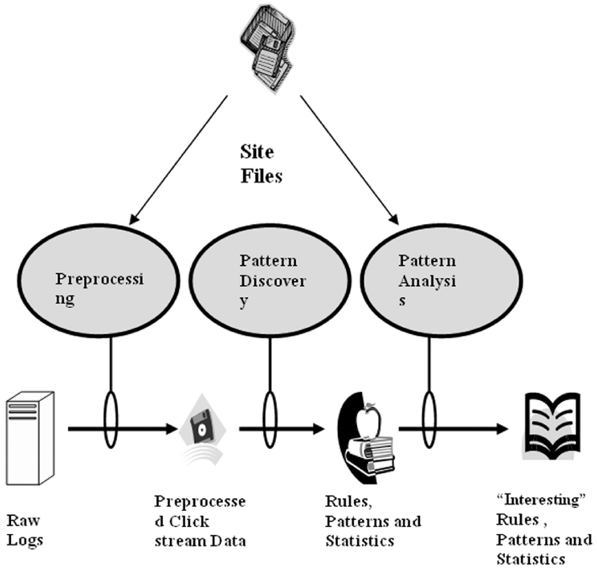
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Click stream Data
Rules, Patterns and Statistics
"Interesting" Rules , Patterns and Statistics

Fig. 1: Web Usage Mining Process

Client Level: It is the client itself which sends information to a repository regarding the users 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 only users whose web clients pass through the proxy. In this paper we will cover only the case of web server (HTTP server) data. The information that we have at the beginning is automatically collected by web 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 it is supported by most of the HTTP servers.

Following is an example line of access log in common log format db01.grohe.it-[19/sep/2001:03:23:53+0100] "GET/HTTP/1.0" 200 4096.

There are various fields in this line:

1. Client IP Address or host name(if DNS look ups are performed);
2. User ID ('-' if anonymous);
3. Access date and time;
4. HTTP request method (GET, POST…);
5. Path of the resource on the web server;
6 Protocol used for the transmission (HTTP|1.0, HTTP |1.1);
7. The status code returned by the server as response(200 for O.K., 404 for not found);
8. The number of bytes transmitted.

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 one, which If are added to the CLF make up the so called combined log format (supported by Apache Web Server).

2.2. Techniques for Web usage Mining

A number of techniques [3] 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 etc. Association Rules [6] consider every URL requested by a user in a visit as item and find out relationships between them with a minimum support level. Sequential Patterns [7] are used to discover time ordered sequence of URL's followed by past users in order to pridict future ones. Clustering [8, 9] forms meaningful clusters of URL's by discovering similar attributes between them according to user behavior.

2.3. Applications of Web Usage Mining (Figure 2)

2.3.1. Personalization

Personalization for a user is achieved by keeping track of previously accessed pages e.g. individualized marketing for E-Commerce[10].Making dynamic recommendations to a web server on the basis of her/his profile in addition to usage behavior is very attractive to many applications e.g. cross sales and up sales in E-Commerce[11].

Web usage mining is an excellent approach for achieving this objective as described in [12] existing recommendation systems.

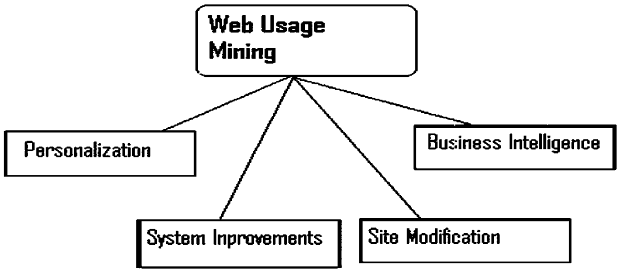

Fig. 2: Applications of Web Usage Mining

2.3.2. System Improvement

Performance and other service quality factors are crucial to user satisfaction from services like databases, networks etc.

Similar qualities can be expected from the user of web services. Web usage mining could provide the key to understand web traffic behavior, which can in term be used for developing policies for web coaching, Network Transmission [14].

2.3.3. Site Modification

The attractiveness of a web site, in term of both content and structure, is crucial to many applications e.g. a 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 [13,14] structure of a Website changes automatically on the basis of usage patterns discovered from server logs.

2.3.4. Business Intelligence

Information on how customers are using a website is crucial for marketers of e-tailing businesses. Buchner et. al [15] have presented a knowledge discovery process to discover marketing intelligence from web data. They defined a web log data hyper cube that would combine web usage data with marketing data for e-commerce applications. 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.

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 [16]. 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 to mine for knowledge. In the following subsections we discuss about the data cleaning and data reduction algorithms.

3.1. Data Cleaning

Data cleaning concerned with removing all the data tracked in web logs that are useless for mining purposes[17,18] e.g. requests for graphical page content(e.g. JPG,.GIF, and css); Request for any other file which might be included in to web page ; or even navigation sessions performed by robots and web spiders. Robots and web spider navigation patterns must be explicitly identified. This is usually done for instance by referring to the remote host name, by referring to the user agent or by checking the access to the robots.txt file. However some robots actually send a false user agent in HTTP request. In these cases, a heuristic based on navigational behavior can be used to separates robot sessions from actual users sessions [19, 20]. An algorithm [21] for cleaning the entries of server logs is presented below -

```
Read record in database.
For each record in database
Read fields (URI – stem) //URI- stem indicates
 The target URL//
If fields = {*.gif,*.jpg,*.css} then
   Remove records
Else
Save records
End if
Next record
```

3.2. Data Reduction

Access log files on the server side consists log information of user who opened a session. These logs include the list of items that a user agent has accessed. The log format of the file is CERN (Common Log Format) which include special record formats. The information in this record is sufficient to obtain session information. In this context, the information which is going to be extracted is defined as the click- stream in a user session for a particular web server. A click stream is defined as a series of page view requests. The parser[22] transforms a set of logs, L into a set of sessions

$$L = \{L_1, L_2, L_3 \ldots\ldots.. L_i, \ldots L|L|\}\ \forall_i <= |L|$$

|L|: The number of records in the log file.

$L\{IP_i, TIME_i, METHOD_i, URL_i, PROT_i, CODE_i, BYTE_i\}$

$S = \{S1, S2, S3,\ldots Si\ldots. S|S|\}, \forall i <= |S|$

|S|: the number of discrete sessions

Each session Si contains $IP_{i,}$ $PAGES_i$ which are navigated in that session

$$S_i = \{IP_i, PAGES_i\}$$

$$PAGES_i = \{URL_{i-1}, URL_{i-2}, \ldots\ldots URL_{i-k}\}$$

Each URL is represented by a vertex in graph model; K is the number of pages which is requested by user agent connected from IPi in session Si.

The set of URLs which is forming a session should satisfy the requirement that the time of elapsed between two consecutive requests is smaller than a given t, which is accepted as 30 minutes [23]. The algorithm given by the parser [Martinzer] is as follows

```
L : the set of input logs.
|L|: the number of input logs
$\Delta t$ : time interval
S : The set of sessions
|S| : The number of sessions.
Input: L,|L|, $\Delta t$
Output: S, |S|
Function Log-Parser (|L|, L, $\Delta t$)
For each Li of L
If METHOD$_i$ is 'GET' and URL$_i$ is 'WEBPAGE' //1.if
If $\exists$ Sk $\in$ Open_Sessions with IPk = IP$_i$ then //2.if
If ((TIMEi-END_TIMES(Sk) )< $\Delta t$) then // 3.if
Sk = (IPk, PAGEk $_{\cup}$ URL$_i$)
Else
   CLOSE_SESSION(Sk)
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
```

In above mentioned algorithm maximum time detected in the page set of corresponding session is returned by the Function END_TIME. Function CLOSE_SESSION is removing corresponding session from OPEN_SESSION set. Function OPEN_SESSION is adding corresponding session to open session set. When an URL$_i$ is being considered, the algorithm examines whether the given URL is invalid format. If any incomplete or invalid format URL is found the algorithm discards the corresponding log entry. Clearly the parser above transforms data definition of logs containing 7 attributes tuple into another data format, Si=(IPi, PAGE$_i$). Data cleaning is performed in URL$_i$->PAGES$_i$ conversion. After that each session is reduced to Si=(PAGE$_i$) format. We have only a sequential list of pages that were visited in sessions. Thus after data reduction we have data in reduced format.

4. Related Work

In this section we present the main related works in this particular area. In the recent years, there has been much research on web usage mining. However, data preprocessing has received far less attention then it deserves. Methods for user identification, session zing, page view identification, path completion and episode identification are presented in [6].In some other work [24], the authors compared time based and referrer based heuristics for visits reconstruction. In [25], 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.

5. Conclusion

In this paper we survey some data preprocessing activities like data cleaning and data reduction. In section 3 we presented the algorithms for data cleaning and data reduction. It is important to note that before applying data mining techniques to discover user access patterns from web log, data must be processed because quality of results is based on data to be mined. In section 2 we presented the overview of web usage mining including techniques and applications. Web usage mining is still new area for research and related issues need further explanation and research.

References

[1] O.Eizoni. The World Wide Web: Quagmire or Gold Mine. Communications of the ACM, 39 CII): 65-68, 1996.

[2] Robert Cooly, Bamshad Mobasher, Jaideep Srivastava (1997) : Web Mining Information and Pattern Discovery on the World Wide Web.

[3] Robert Cooly, Bamshad Mobasher, Jaideep Srivastava (1999) : Data Preparation for Mining World Wide Web browsing Pattern.

[4] http://www.w3.org/Daemon/user/config/logging.html # common - log – file -format.

[5] Jaideep Srivastva, Robert Cooly, Mukand Deepande, Pang-MingTan (2000) : Web Usage Mining : Discovery and Applications of usage Patterns from Web Data.

[6] Karuna P. Joshi,Anupam Joshi and Yelena Yesha: on using a Ware-house to Analyze Web Logs. Distributed and Parallel Databases, 13(2):161-180, 2003.

[7] Eleni Stroulea Nan niu and Mohammad El-Ramly. Understanding Web Usage for Dynamic Web Site Adaptation: A Case Study. In Proceedings of Fourth International Workshop on Web Site Evolution (WES, 02). Pages 53-64, IEEE, 2002.

[8] A. Banarjee and J.Ghosh. "Clickstream Clustering using Weighted Longest Common Subsequences". In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 2001.

[9] Jeffery Heer and Edti. Chi. "Mining the Structure of User Activity using Cluster Stability". In Proceeding of the Workshop on Web Analytics, Second SIAM Conference on Data Mining. ACM press, 2002.

[10] Broadvision http://www.broadvision.com.

[11] Bamshad Mobashar, Robert Cooly, Jaideep Srivastava. Creating Adaptive Web Sites Through usage Based Clustering of URLs in Knowledge and Data Engineering Workshop 1999.

[12] E. Cohen, B. Krishnamurthy and J. Rexford. "Improving end to End Performance of the Web using Server Volumes and Proxy Filters". In Proc. ACM SIGCOMM pages 241-253, 1998.

[13] Mike Perkowitz and Oren Etzioni. Adaptive Web Sites : Automatically Synthesizing Web Pages. In Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.

[14] Peter Pirolli, James Pitkow and Ramna Rao.Silk from and Sow's Ear : Extracting usable Structure from the Web. In CHI – 96 Vancouver, 1996.

[15] Alex Buchner and Maurice D Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. SIGMOD record, 27(4) : 54-61, 1998.

[16] David A. Grossman and Ophir Frieder. Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series) (2nd Edition) (Paperback - Dec 20, 2004).

[17] Boris Diebold and Michael Kaufmann. usage based Visualization of Web Localities. In Australian Symposium on Information Visualization Pages 159-164, 2001.

[18] Corin R. Anderson: A Machine Learning Approach to Web Personalization Ph. D. Thesis, University of Washington, 2002.

[19] 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.

[20] Pang-Ning Tan and Vipin Kumar. Discovery of Web Robot Sessions based on their Navigational Patterns. Data Mining and Knowledge Discovery, 6(1) : 9-35, 2002.

[21] Mohd Helmy, Abd Wahab, Nik Shahidah. Development of Web usage Mining Tools to Analyze the Web Server Logs using Artificial Intelligence Techniques. The 2nd National Intelligence Systems and Information Technology Symposium (ISITS 207), Oct 30-31, 2007, ITMA-UPM, Malaysia.

[22] Martinez E. Karamcheti V. "A Prediction Model for User Access Sequence" In WEBKDD Workshop : Web Mining for usage Patterns and user Profile, July 2002.

[23] Catledg, L; Pitkow, J.: "Characterizing Browsing Behaviors on the World Wide Web", In Computer Networks and ISDN System 27(E) 1995.

[24] 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.

[25] 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.