Customized Web Log Preprocessing through Web Mining

Ashoksinh V. Solanki
I/C Principal College of Applied Sciences & Professional Studies, Chikhli
(BCA Programme)
Affiliated to Veer Narmad South Gujarat University, Surat(Gujarat)-India

Abstract

The World Wide Web (WWW) continues and consistence to grow in both size and complexity. Every click or transaction users do is valuable data for website owners especially for e-business companies. This data is kept in data warehouses for future use. If we call the Internet as a laboratory then data warehouses and databases are the places to put data and data miners are experts who perform some experiments and find some new information. More clearly data miners evaluate and filter as a result convert data to information and information to knowledge by performing some techniques.

The contribution of this paper is towards the various areas containing web sites on internet, which can make best use of different web mining techniques to improve their business decisions based on the user behavior analysis which can ultimately help in improving the relevance of their web site to suit their user needs and adding value to their business growth. It also contributes about the factors responsible and governing the usage of web mining for the web sites to improve business intelligence. This paper is main emphasis on customized web log preprocessing techniques which will help us to save time and we identify user behavior.

KEYWORDS: Web Mining, Web Usage Mining, Web Logs, Web Log preprocessing

1. INTRODUCTION

Web Data mining applications are widely used in various areas such as e-business, education, telecommunications, financial, engineering, biotechnology etc. Web mining is used for most of these applications. While Web mining has great benefits and it also has great challenges.

Most of the data on the web is either unstructured or semi-structured. Before mining data should be prepared and reconstructed to a new form. The information on the internet is in the form of static and dynamic web pages of various areas from education, industry to every walk of life including blogs.

Web sites are having inter, intra linked web pages. The speed of increase of web information is rapid. The hidden knowledge discovery, patterns and trends of user access can be found from the way the web sites and web pages are accessed and it is useful from the business perspective giving future directions for decision making.

The Data Mining techniques help in identifying the patterns implying the future trends in the studied data. The Web Mining is an application of the data mining techniques to find interesting and potentially useful Knowledge from web data.

1.1.Why researchers use web data mining?
The Various Business Areas where Web Mining has helped in Improving the Business Decision Making

1.1.1 E-Business
Analysis of click-stream data i.e. web mining uncovers real-time e-business opportunities across geography.
It provides ways to target right customers and understand their needs and to customize services and strategies in near-or-real time. The area of advertising is no exception for utilizing the opportunities provided by online customer analytics to promote right products in real time to the right customer. It also helps in effectiveness of a web site as a channel for marketing by quantifying the user's behavior while on the web site.

1.1.2. CRM
Analytical CRM utilizes business intelligence and reporting methodologies such as data mining and analytical processing to CRM applications. While the earlier CRM implementations focus on improving operational efficiencies in the sales and service functions through tailor-made solutions for call-center Management. With the amount of available online content, today organizations put premium on understanding, adopting and managing the same, convert them into appropriate knowledge suitable to serve their customers better, and thus improve the operations and accelerate the process of delivery of products to markets.

1.1.3. Customer behavior
Web Mining helps in understanding the concerns such as current and future probability of every customer, relationship between behavior and the loyalty at the website. Web data mining also used for predicting customer's future behavior that is essential for website content planning and design.

1.1.4 Web Usage Mining For Proxy Server
Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years. Commercial companies as well as academic researchers have developed an extension array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular website. Performing this kind of investigations on your website can provide information that can be used to better accommodate the user's needs.

2.0. WEB DATA MINING TECHNIQUES
Web Mining techniques make use of the web information and are based on web content mining, web structure mining, and web usage mining. Web data is Web content have text, image, records, etc. Web structures have hyperlinks, tags, etc. and Web usage can have http logs, app server logs, etc.

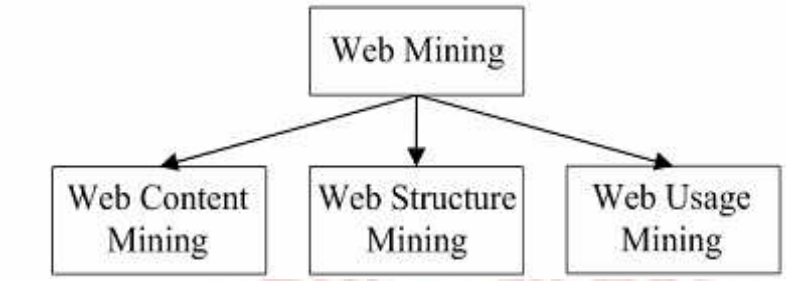

Figure 1 : Taxonomy of Web Mining

2.2.1 Web Content Mining

It deals with discovering useful information or knowledge from web page contents than hyperlinks and goes beyond using keywords in a search engine. Web content consists of information such as unstructured free text, image, audio, video, metadata, and hyperlink. Search engines, subject directories, intelligent agents, cluster analysis, and portals are used to find out what a user might be looking for.

2.2.2 Web Structure Mining

It deals with discovering and modeling the hyperlink structure of the web pages based on the topology of the hyperlinks. This gives similarity between sites or sites for a particular topic or web communities.

2.2.3 Web Usage Mining

It deals with understanding user behavior with a web site and to obtain information that may assist in web site reorganization to suit user needs. The mined data includes data logs of user's web interaction, having web server logs, proxy server logs and browser logs, having data about referring page, user identification, user spent time at site and sequence of pages visited and also cookie files that contain information.

Web usage mining identifies a complex procedure by which statistical methods and data mining technologies are employed in order to extract implicit, previously unknown and potentially useful information from web data.

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web.

Concept of web usage mining

Discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers.

Typical Sources of Data:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies
2. E-commerce and product-oriented user events

(e.g. shopping cart changes, add or delete product Click-through, etc.)

3. User profiles and/or user ratings

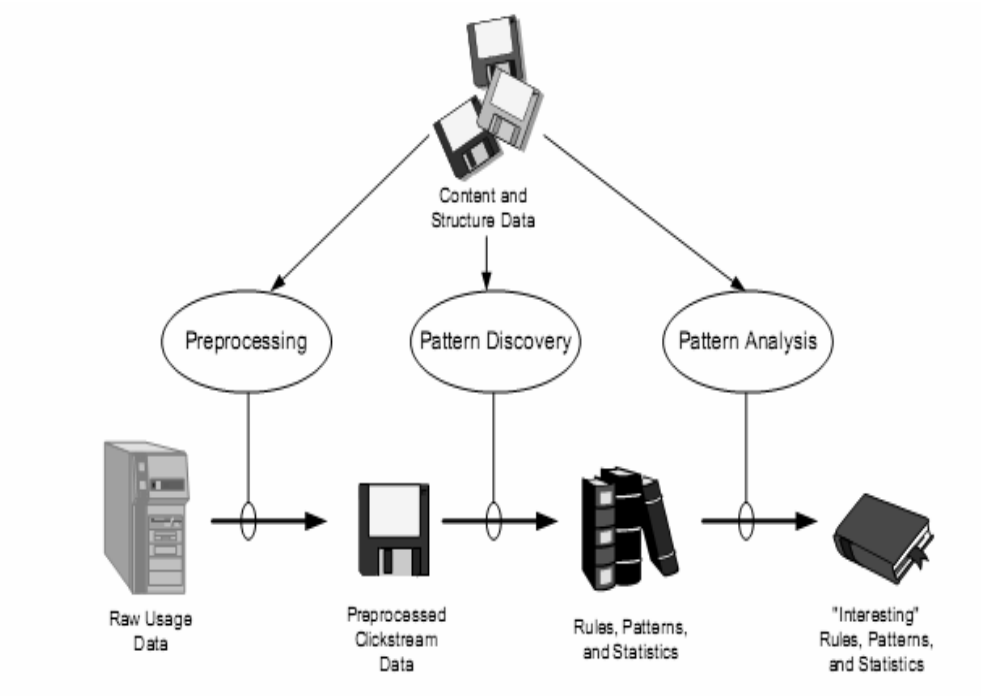

Figure 2 : Web Usage Mining Process

Web Log Format

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format as shown in Figure 3.

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

Figure 3 : Common Web Log Format

Approach of Web usage mining

The web usage mining generally includes the following

Several steps: data collection, data pretreatment, knowledge discovery and pattern analysis.

Data collection:

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must be used scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data.

Data pretreatment:

Some databases are insufficient, inconsistent and including noise. The data pretreatment 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. In the data pretreatment work, mainly include data cleaning, user identification, session identification and path completion.

1) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed:

1. The records of graphics, videos and the format information. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record;

2. The records with the failed HTTP status code by examining the Status field of every record in the web access log. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

2) User and Session Identification:

The task of user and session identification is find out the different user sessions from the original web access log. User's identification is to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

a. The different IP addresses distinguish different users;

b. If the IP addresses are same, the different browsers and operation systems indicate different users.

c. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URL field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty.

d. The session identified may contain more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

3) Path completion

Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. With the help of data pretreatment, web log can be transformed into another data structure, which is easy to be mined.

Knowledge Discovery

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions. Use statistical method to carry on the analysis and mine the pretreated data.

Pattern analysis

Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. In pattern analysis first discard the less significance rules or models from the interested model storehouse; Next use technology of OLAP(On-Line Analytic Process) and so on to carry on the comprehensive mining and analysis and final step to let discovered data or knowledge be visible. In this work pattern discovery means applying the introduced frequent pattern discovery methods to the log data. For this reason the data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of the algorithms. Log files are stored on the server side, on the client side and on the proxy servers. Logs are processed in Common Log Format. Pattern analysis means understanding the results obtained by the algorithms and drawing conclusions. In pattern discovery phase methods and algorithms used have been developed from several fields such as statistics, machine learning, and databases.

3. WEB LOG PREPROCESSING

The inputs to the preprocessing phase are the log and site files. The outputs are the user session file and transaction file. Web servers register a Web log entry for every single access they get, in which important pieces of information about accessing are recorded, including the URL requested, the IP address from which the request originated, and a timestamp. A log file can be located in three different places-Web Servers, Web proxy Servers, and Client browsers. Preprocessing contains four sub steps: Data Cleaning, User Identification, Session Identification and Formatting.

Pre-processing method gives output for some particular application, but different web application requires different pre-processing of logs. Multimedia application requires log of multimedia link request like log having jpg, mpg, and gif etc. resource. E-commerce

application requires different user requests. Traditionally web log pre-processing removes http error log, css access log etc. We introduced one more step in traditional pre-processing steps, before data cleaning is Customization. In this step we clean log on the basis of user requirement for application. User selects choice of application for which he wants to perform usage mining according to that our pre-processing approach works.

Steps for customized web log pre-processing are as follows:
1. Get Input as raw web accessing log file.
2. User choice for normal, multimedia, graphics or E-Commerce applications.
3. Read raw web log file and remove logs according to user selection to generate intermediate file.
4. Identify users and their resources uniquely and assign a unique ID.
5. Identify resource accessed by users according to ID.
6. Create preprocessed file by mapping of user ID and Resource IDs accessed by them.

With customized web log pre-processing we can reduces file size for pattern analysis according to application which reduces time to find frequent pattern or user behavior.

4. CONCLUSION

Research work includes preprocessing phase of web usage mining which can be utilized in industry and application oriented system. We uses customized web log preprocessing rather than traditional approach which may reduces size of raw web log file. In future research, work is carried out for developing a web usage mining tool with customized web log preprocessing and combined pattern analysis approaches according to different application.

5. ACKNOWLEDGMENT

I wish to acknowledge the encouragement of my entire CASPS family which helps me to work hard towards producing this work.

6. REFERENCES

Data Mining - Typical Data mining Process for Predictive Modeling-BPB Publications First Edition 2004 REPRINTED 2007

Mining the WWW – An information search approach - George Chang

Web Data Mining – exploring hyperlinks, contents and usage data -Bing Liu Second Edition July 2011

Web Usage mining - Bamshad Mobasher

Data Mining and Knowledge Discovery Techniques - Dr. Carlo kopp

http://www.jafsoft.com Web server log

http://www.statsoft.com