A Critique on Web Usage Mining

K.S.R. Pavan Kumar V.V. Sreedhar L. Manoj Chowdary

Abstract:Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs.

A complete review on web usage mining is discussed in this paper. The life style and the study of the great changes taking place, much higher efficiency, resource of information are the greatest degree of sharing are increasing day to day because of the increasing popularity of the internet through web access to more information and data. Web is a combination of Web mining technology and Data mining and this is an integrated technology resources extracted from WWW information of the course is the implication of the web resource. Web mining attempts to determine useful information and knowledge from secondary data obtained from the user interactions with the web. Data mining involves the study of data-driven techniques to discover and model hidden patterns in large volumes of raw data. The application of data mining techniques to Web data is referred to as Web data mining. Researchers have identified three broad categories of web mining: web content mining, web structure mining and web usage mining.

The knowledge discovery and the data mining methods usage on the web is now on the public eye of a boosting number of researchers. Web usage mining is a kind of data mining method takes the help of users' session and behavior and recommends the web usage patterns. In Web usage mining there are three processes which comes into consideration, namely, preprocessing, pattern discovery and pattern analysis. In web usage mining there are different already existing techniques. There are different advantages and disadvantages for those existing techniques. This paper gives a sight on some of the existing web usage mining techniques.

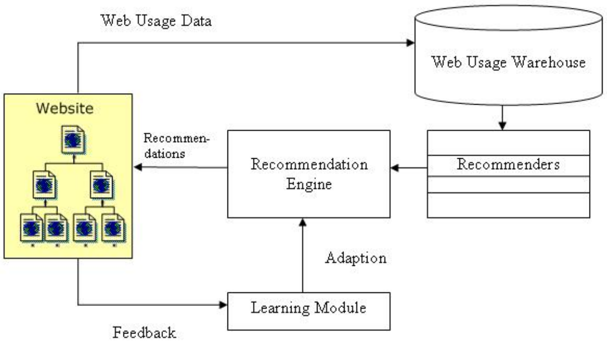
INTRODUCTION:

Web mining has become very vital for effective web site personalization and management. It is crucial for network traffic flow analysis, creating business services, business support, etc. Web mining can be classified into three different types, which are Web content mining, Web structure mining and Web usage mining.

Web content mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. It is the application of one of the data mining techniques to content published on the internet, usually as semi structured, un-structured, structured documents. The most widely studied research topics of Web content mining is structured data extraction. Structured data on the Web are often very important as they represent their host pages essential information. The extraction of such data allows us to provide different value added services like meta-search and shopping. The studies made by researchers in AI and data mining and database reveals the problem that in contrast to unstructured texts, structured data is also easier to extract.

Web structure mining: Web Structure Mining can be is the process of discovering structure information from the Web. Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics such as PageRank. Web's hyperlink structure is operated by the Web structure mining. The illustration of the information about pages ranking or authoritativeness and enhance search results through filtering is provided by this graph structure. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level. The research at the hyperlink level is also called Hyperlink Analysis.

Web usage mining: User interaction with a web server, including web logs, click streams, and database transactions results at a web site or a group of a related sites is analyzed by Web usage mining. Analyzing the web usage log data web mining systems can discover knowledge about users' interest and systems usage characteristics. Personalization and collaboration in decision support, website design, website evaluation, marketing and web based systems are the various applications of such knowledge. For effective web site management, creating adaptive web sites, business and support services, personalization and network traffic flow analysis web usage mining has become very crucial.



Web Usage Mining is a part of web mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage mining involves mining the usage characteristics of the users of Web Applications and this extracted information can then be used in a different ways such as checking of fraudulent elements and improvement of the application.

This paper provides a survey and analysis of current Web usage mining technologies and systems. A Web usage mining system must be able to perform five major functions:

i. Data Gathering,
ii. Data Preparation,
iii. Pattern discovery,
iv. Pattern analysis, and
v. Pattern Applications.

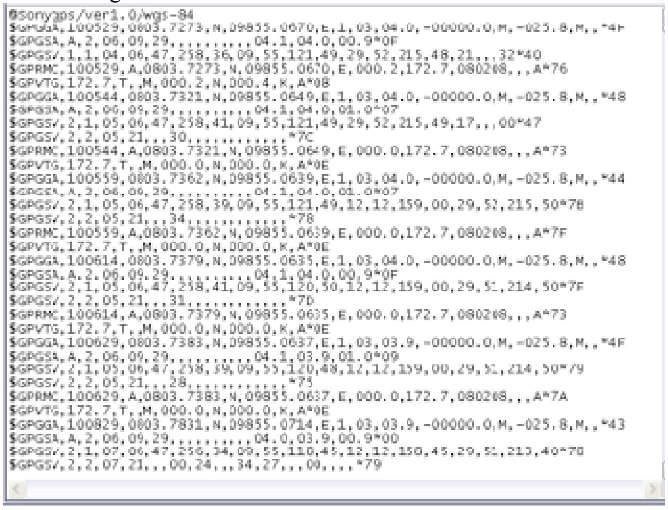
1. DATA GATHERING

The usage data collected at different sources represent the navigation patterns of different segments of the overall web traffic, ranging from single user, single site browsing Behavior to multi-user, multi-site access patterns. The Server log files are the primary data sources in web usages mining, which includes Web server access logs and

Application server logs. Web server log does not accurately contain sufficient information for inferring the behavior at the client side as they relate to the pages served by the web server. A Web log file records activity information whenever a user requests a Web server. A log file can be located in three different places:

i) Web servers,
ii) Web proxy servers, and
iii) Client browsers

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Examples of the types of information the server preserves include the user's domain, subdomain, and hostname; the resources the user requested the time of the request; and any errors returned by the server. Each log provides different and various information about the Web server and its usage data. Most logs use the format of a common log file or extended log file



```
@sonygps/ver1.0/wgs-84
$GPGGA,100529,0803.7273,N,09855.0670,E,1,03,04.0,-00000.0,M,-025.8,M,,*4F
$GPGSA,A,2,06,09,29,,,,,,,,,,04.1,04.0,00.9*0F
$GPGSV,1,1,04,06,47,258,36,09,55,121,49,29,52,215,48,21,,,32*40
$GPRMC,100529,A,0803.7273,N,09855.0670,E,000.2,172.7,080208,,,A*76
$GPVTG,172.7,T,,M,000.2,N,000.4,K,A*08
$GPGGA,100544,0803.7321,N,09855.0649,E,1,03,04.0,-00000.0,M,-025.8,M,,*48
$GPGSA,A,2,06,09,29,,,,,,,,,,04.1,04.0,01.0*07
$GPGSV,2,1,05,06,47,258,41,09,55,121,49,29,52,215,49,17,,,00*47
$GPGSV,2,2,05,21,,,30,,,,,,,,,,,,*7C
$GPRMC,100544,A,0803.7321,N,09855.0649,E,000.0,172.7,080208,,,A*73
$GPVTG,172.7,T,,M,000.0,N,000.0,K,A*0E
$GPGGA,100559,0803.7362,N,09855.0639,E,1,03,04.0,-00000.0,M,-025.8,M,,*44
$GPGSA,A,2,06,09,29,,,,,,,,,,04.1,04.0,01.0*07
$GPGSV,2,1,05,06,47,258,39,09,55,121,49,12,12,159,00,29,52,215,50*78
$GPGSV,2,2,05,21,,,34,,,,,,,,,,,,*78
$GPRMC,100559,A,0803.7362,N,09855.0639,E,000.0,172.7,080208,,,A*7F
$GPVTG,172.7,T,,M,000.0,N,000.0,K,A*0E
$GPGGA,100614,0803.7379,N,09855.0635,E,1,03,04.0,-00000.0,M,-025.8,M,,*48
$GPGSA,A,2,06,09,29,,,,,,,,,,04.1,04.0,00.9*0F
$GPGSV,2,1,05,06,47,258,41,09,55,120,50,12,12,159,00,29,51,214,50*7F
$GPGSV,2,2,05,21,,,31,,,,,,,,,,,,*7D
$GPRMC,100614,A,0803.7379,N,09855.0635,E,000.0,172.7,080208,,,A*73
$GPVTG,172.7,T,,M,000.0,N,000.0,K,A*0E
$GPGGA,100629,0803.7383,N,09855.0637,E,1,03,03.9,-00000.0,M,-025.8,M,,*4F
$GPGSA,A,2,06,09,29,,,,,,,,,,04.1,03.9,01.0*09
$GPGSV,2,1,05,06,47,258,39,09,55,120,48,12,12,159,00,29,51,214,50*79
$GPGSV,2,2,05,21,,,28,,,,,,,,,,,,*75
$GPRMC,100629,A,0803.7383,N,09855.0637,E,000.0,172.7,080208,,,A*7A
$GPVTG,172.7,T,,M,000.0,N,000.0,K,A*0E
$GPGGA,100829,0803.7831,N,09855.0714,E,1,03,03.9,-00000.0,M,-025.8,M,,*43
$GPGSA,A,2,06,09,29,,,,,,,,,,04.0,03.9,00.9*00
$GPGSV,2,1,07,06,47,256,34,09,55,110,45,12,12,150,45,29,51,213,40*70
$GPGSV,2,2,07,21,,,00,24,,,34,27,,,00,,,,*79
```

Fig: Server Log File

The content data in a site is the collection of objects and relationships that is conveyed to the user. The structure data represents the designer's view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. The operational database for the site may include additional user profile information. Web usage data are usually supplied by two sources: trial runs by humans and Web logs. The first approach is impractical and rarely used because of the nature of its high time and expense costs and its bias. Most usage mining systems use log data as their data source. This section looks at how and what usage data can be collected

Databases are used instead of simple log files to store information so to improve querying of massive log repositories. Internet service providers use proxy server services to improve navigation speed through caching. Collecting navigation data at the proxy level is basically the same as collecting data at the server level but the proxy servers collects data of groups of users accessing groups of web servers. Usage data can be tracked also on the client side by using JavaScript, Java applets, or even modified browsers

2. Data Pre-Processing or Data Preparation:

Data pre-processing is often neglected but an authoritative step in the data mining process. Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. In a customer relationship management (CRM) context, data preprocessing is a component of Web mining. Web usage logs may be preprocessed to extract meaningful sets of data called user transactions, which consist of groups of URL references. User sessions may be tracked to identify the user, the requested websites and their order, and the amount of time spent on each one. As soon as when these have been differentiated from the raw data, they yield more information such as marketing, consumer research and personalization.

There are different number of methods and tools used for pre-processing. These methods and tools also include sampling, feature extraction, denoising, normalization and transformation. Sampling method means selecting a representative subset from a large population of data. Feature extraction, means pulling out specific data that is significant in some particular context. Denoising, means removing noise from data. Normalization is a method which organizes data for more effective and efficient access. Transformation is a method which manipulates raw data to produce a single input. Pre-processing technique is also useful for association rules algorithms. The raw web log data after pre-processing and cleaning could be used for pattern discovery, pattern analysis, web usage statistics, and generating association/ sequential rules. Much work has been performed on extracting various pattern information from web logs and the application of the discovered knowledge range from improving the design and structure of a web site to enabling business organizations to function more efficiently .Data pre-processing involves mundane tasks such as merging multiple server logs into a central location and parsing the log into data fields.

The preprocessing comprises of

1. Data cleaning which consists of removing all the data tracked in Web logs that are useless for mining purposes.
2. The reconstruction and identification of the users" sessions, and
3. Data formatting.

Problem with huge real-world database

1. Incomplete Data:-
   I. Noisy
   II. Inconsistent
   III. Missing value
2. Major Tasks in Data Pre-processing
   I. Data cleaning
   II. Data integration
   III. Data transformation
   IV. Data reduction

a. *Data Cleaning*

Data cleaning is usually site-specific, and involves tasks such as, removing extraneous references to embedded objects that may not be important for the purpose of analysis, including references to style files, graphics, or sound files. The cleaning process also may involve the removal of at least some of the data fields. The status code return by the server is three digit numbers. There are four class of status code: Success (200 Series), Redirect (300 Series), Failure (400 Series), Server Error (SOD Series). The most common failure codes are 401 (failed authentication), 403 (Forbidden request to a restrict subdirectory, and the dreaded 404 (file not found) messages.

Such entries are useless for analysis process and therefore they are cleaned form the log files.

b. Data Integration

After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The clean server log can be thought of in two ways; either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. The goal of transaction identification is to create meaningful clusters of references for each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. This process can be extended into multiple steps of merge or divide in order to create transactions appropriate for a given data mining task. A transaction identification module can be defined as either a merge or a divide module. Both types of modules take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the module in the same format as the input. The requirement that the input and output transaction format match allows any number of modules to be combined in any order, as the data analyst sees fit.

c. Data Conversion

The log file is simple text file contain various parameters like client IP address, client name, date, time, instant name, server name, server IP, status codes, method and page name. Web server generate a log entry for every page (hyper link clicked by user) viewed by user.

The TransLog algorithm convert such log file into Access table or Oracle table which is further useful for data mining and other action. The TransLog algorithm gives the actionable data. It transform the data contain in simple text file to table. Apart from the TransLog algorithm also there the Web Log analyzers which are used to retrieve the data patterns from the Server Log files.

d. Data Reduction

Data reduction is the transformation of numerical or alphabetical digital information derived empirical or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts. When information is derived from instrument readings there may also be a transformation from analog to digital form. When the data are already in digital form the 'reduction' of the data typically involves some editing, scaling, coding, sorting, collating, and producing tabular summaries. When the observations are discrete but the underlying phenomenon is continuous then smoothing and interpolation are often needed. Often the data reduction is undertaken in the presence of reading or measurement errors. Some idea of the nature of these errors is needed before the most likely value may be determined.

3) PATTERN DISCOVERY

One of the major goals of Web usage mining is to reveal interesting trends and patterns. Such patterns and statistics can often provide important knowledge about a company's customers or the users of a system. In pattern discovery and analysis, generic machine learning and data mining techniques, such as association rule mining, classification, and clustering, can often be applied.

A. Clustering:

Clustering techniques work by identifying groups of consumers who appear to have similar preferences. Once the clusters are created, averaging the opinions of the other consumers in her cluster can be used to make predictions for an individual. Some clustering techniques represent each user with partial participation in several clusters. The prediction is then an average across the clusters, weighted by degree of participation.

B. Classification:

Classifiers are general computational models for assigning a category to an input. The inputs may be vectors of features for the items being classified or data about relationships among the items. The categories are a domain specific classification such as malignant/benign for tumor classification, approve/reject for credit requests or intruder/authorized for security checks. One way to build a recommender system using a classifier is to use information about a product and a customer as the input, and to have the output.

C. Association Rules Mining:

Association rule mining is to search for interesting relationships between items by finding items frequently appeared together in the transaction database. If item B appeared frequently when item A appeared, then an association rule is denoted as A -confidence are two measures of rule interestingness that reflect usefulness and certainty of a rule respectively. Support, as usefulness of a rule, describes the proportion of transactions that contain both items A and B, and confidence, as validity of a rule, describes the proportion of transactions containing item B among the transactions containing item A. The association rules that satisfy user specified minimum support threshold (minSup) and minimum confidence threshold (minCon) are called Strong association rules. One of the best-known examples of web mining in recommender systems is the discovery of association rules, or item-to-item correlations. Association rules have been used for many years in merchandising, both to analyze patterns of preference across products, and to recommend products to consumers based on other products they have selected.

4) PATTERN ANALYSIS

Pattern Analysis is a final stage of the whole Web usage mining. The main motto of this process is to remove irrelevant patterns rules and to extract the interesting patterns or rules from the output of the pattern discovery process. The output of earlier stage web usage mining is often not suitable for the web site administrators. The type of information sought in this respect is "How people are using the site? Which of them are accessed most frequently?" These queries require analysis of the structure of hyperlinks as well as the contents of the page which is done with the help of some analysis tools and methodologies. Most common used techniques for pattern analysis are OLAP Techniques, Data and Knowledge Querying, Data and Knowledge Querying, Visualization Techniques and Usability Analysis.

OLAP Techniques

On-line Analytical Processing (OLAP) is emerging as a very powerful paradigm for strategic analysis of databases in business settings. Some of the key characteristics of strategic analysis include: very large data volume, explicit support for the temporal dimension, support for various kinds of information aggregation, and long-range analysis in which overall trends are more important than details of individual data items.

Data and Knowledge Querying

One of the reasons attributed to the great success of relational database technology has been the existence of a high-level, declarative, query language, which allows an application to express what conditions must be satisfied by the data it needs, rather than having to specify how to get the required data. Given the large number of patterns that may be mined, there appears to be a definite need for a mechanism to specify the focus of the analysis. First, constraints may be placed on the database to

Restrict the portion of the database from which to mine for. Second, querying may be performed on the knowledge that has been extracted by the mining process, in which case a language for querying knowledge rather than data is needed.

Visualization Techniques

Visualization has been used very successfully in helping people understand various types of phenomena, both real and abstract. Hence it is a natural choice for understanding the behavior of web users. According Groth the visualization is simply the graphical presentation of data.

Usability Analysis

The first step undertaken in this method is to develop instrumentation methods that collect data about software usability. This data is then used to build computerized models and simulations that explain the data. Finally, various data presentation and visualization techniques are used to help an analyst understand the phenomenon. This approach can also be used to model the browsing behavior of users on the web, however, as most of those techniques are disliked by users because of slow speeds, inflexibility, difficult to maintain and limited functionality. To develop a more efficient, flexible and powerful set of tools to undertake this task there still remains a lot of work to be undertaken by both researcher and developer.

5) PATTERN APPLICATIONS:

Web usage mining has been used for various purposes. For example, Buchner and Muhenna proposed a knowledge discovery process for mining marketing intelligence from Web data. Data such as Web traffic patterns also can be extracted from Web usage logs in order to improve the performance of a Web site. Many commercial products have been developed to support analysis and mining of Web site usage and Web log data.

CONCLUSION

An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web usage mining due to the characteristics of click stream data and its relationship to other related data collected from multiple sources and across multiple channels.

The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process, and often requires the use of special algorithms and heuristics not commonly employed in other domains. Web usage mining has emerged as the essential tool for realizing more personalized user-friendly and business optimal Web services. Advances in data pre-processing, modeling, and mining techniques, applied to the Web data,

Have already resulted in many successful applications in adaptive information systems, personalization services, Web analytics tools, and content management systems. As the complexity of Web applications and user's interaction with these applications increases, the need for intelligent analysis of the Web usage data will also continue to grow. Web usage analysis is used to understand the relationship of user and item which exist in the particular sessions However, without the benefit of deeper domain knowledge, such patterns provide little insight into the underlying reasons for which such items or users are grouped together. Thus, a focus on techniques and architectures for more effective integration and mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications, and more sophisticated tools for Web usage mining that can derive intelligence from user transactions on the Web.

REFERENCES

[1]. DYNAMIC MODELLING APPROACH FOR WEB USAGE MINING USING OPEN WEB RESOURCES-International Journal of Engineering Science and Technology (IJEST).

[2]. Xuli Zong, Wen-Chen Hu, Chung-wei Lee-World Wide Web Usage Mining Systems and Technologies.

[3]. An Effective and Complete Preprocessing for Web Usage Mining-International Journal on Computer Science and Engineering (IJCSE).

[4]. A Survey on Web Usage Mining -Global Journal of Computer Science and Technology

[5]. Business Intelligence from Web Usage Mining-Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375{390}

[6]. Web Usage Mining: An Implementation by Aniket Dash and Liju Robin George.

[7]. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data