A Survey on Web Usage Mining, Applications and Tools

ANITHA TALAKOKKULA
Department of Computer Science and Engineering, Stanley College of Engineering and Technology Hyderabad
E-mail: tanitha@stanley.edu.in

Abstract
World Wide Web is a vast collection of unstructured web documents like text, images, audio, video or Multimedia content. As web is growing rapidly with millions of documents, mining the data from the web is a difficult task. To mine various patterns from the web is known as Web mining. Web mining is further classified as content mining, structure mining and web usage mining. Web usage mining is the data mining technique to mine the knowledge of usage of web data from World Wide Web. Web usage mining extracts useful information from various web logs i.e. users usage history. This is useful for better understanding and serve the people for better web applications. Web usage mining not only useful for the people who access the documents from the World Wide Web, but also it useful for many applications like e-commerce to do personalized marketing, e-services, the government agencies to classify threats and fight against terrorism, fraud detection, to identify criminal activities, the companies can establish better customer relationship and can improve their businesses by analyzing the people buying strategies etc. This paper is going to explain in detail about web usage mining and how it is helpful. Web Usage Mining has seen rapid increase towards research and people communities.

1. Introduction
The World Wide Web is an immense source of data that can come either from the web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information daily collected by all the servers around the world. Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. Lots of people started their e-business over the web due to the ease and speed of their business transactions using internet. This requires the study of people personal information, their interests, their economy, and their buying patterns so that their marketing strategies can be improved according to the people and also can produce the goods of their choice of interest. This leads to the web usage mining to find out the various data mining patterns like clustering, classification, association analysis which are required in the e-commerce applications. Data mining applied to the web is known as Web mining. Basically Web mining is divided into three parts: Web content mining, Web Structure Mining and Web Usage mining. Section 2 deals with the classification of web mining. Section 3 deals with the types of patterns mined using data mining. Section 4 deals with the Web usage mining in detail. Section 5 deals with the applications of the Web usage mining. Section 6 deals with the various tools used in web usage mining.

2. Classification of Web mining
With the explosive growth of data available on the World Wide Web, discovery and analysis of useful information from the World Wide Web becomes a practical necessity. Web mining is the application of data mining technologies to huge Web data repositories. Basically, there are three domains that pertain to Web mining: Web content mining Web Structure mining and Web usage mining. The first one is the process of extracting knowledge from the content of Web sites, the second one is the mining the authorized sites based on the links associated with the web sites using the graphical structure, whereas the last one is also known as Web log mining, is the process of extracting interesting patterns from Web access logs.

2.1 Web content mining
Web Content Mining is that part of Web Mining which focuses on the raw information available in Web pages. Source data mainly consists of textual data in Web pages (e.g., words, but also tags). Typical applications of web content mining are content-based categorization and content-based ranking of Web pages. The basic purpose of this is mining, extraction and integration of useful data, information and knowledge from Web page contents.

2.2 Web structure mining
Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Nodes of a graph are web pages and lines are hyper links. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.
2.3 Web usage mining

Web usage mining is the process of extracting useful information from various web logs i.e. users history. Various web logs are server logs, client logs and network logs. This is the process of finding out what users are looking for on the Internet and usage of web pages. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

3. Various Patterns of Data Mining

Data mining is the process of extracting various patterns from the large data sources like databases, data warehouses, advanced databases. The various steps takes place during data mining are: data cleaning, data integration, data transformation, pattern evaluation and visualization. During data cleaning process, it removes the noise and inconsistencies which may misleads the knowledge extracted. Data Integration, collects the data from the various data sources into the single data cube before data mining takes place. Data transformation transforms the data into unified format before storing into the data cubes, this is needed because the data may be in different formats/ different ranges for different data sources. Pattern- Evaluation allows user to select the choice of pattern to be evaluated during data mining. There are various patterns like classification, clustering, association analysis, sequential analysis, characterization, Discrimination etc. Finally Visualization step involves the various ways of visualizing the patterns extracted from the data mining like graph, pie chart, bar graph, data cubes, statistics etc.

Classification is the one of the pattern is used to classify the data into various pre determined groups. This is basically uses training set to group the data. In the Web domain classes usually represent different user profiles and classification is performed using selected features that describe each user’s category. Mostly used classification algorithms are decision tree, Naïve-Bayes and neural network’s back-propagation.

Clustering is used to group the items together that have similar characteristics. In the context of Web mining, we can distinguish two cases, user clusters and page clusters. Page clustering identifies groups of pages that seem to be conceptually related according to the users’ perception. User clustering results in groups of users that seem to behave similarly when navigating through a web site. Various clustering algorithms which are used most are k-means based on the distance, CURE based on the hierarchical structuring, DBSCAN based on the density, COBWEB based on the model etc.

Association rule mining is a technique for finding frequent patterns, associations, and correlations among sets of items. Association rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Association analysis is the important pattern used in e-commerce application which is used basically for analyzing the buying behaviour of the people. This express the buying behaviour in the form of association rules such as Milk ^ Bread => Eggs. This helps the marketing strategies for improving their business, to establish the customer relationships, organization of items in the shop, helps to analyze which items to put in discounts and which items to produce together and in which season which items are moving more etc. Mostly used algorithm for finding association rules is Apriori-algorithm.

Sequential pattern discovery is an extension of association rules mining in that it reveals patterns of co-occurrence incorporating the notion of time sequence. In the Web domain such a pattern might be a Web page or a set of pages accessed immediately after another set of pages. This analyzes which items are moving together in certain time period or after one item is purchased when is the other related item customer is going to purchase etc. For example after purchasing computer, with-in a week customer is going to purchase software.

Characterization is the other pattern that can be evaluated in data mining. This is basically summarizing the data with their various statistics that helps to understand and analyzing for the future use.

Discrimination is similar to Characterization to summarize the data, but this compares the target data and produces the comparative study. This is mostly used in the marketing when we are comparing two products data or comparing two companies for improving the business.

4. Web Usage Mining

4.1 Introduction

Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers, web clients and proxy servers. Web usage mining is the process of extracting useful information from various web logs to find which of the pages are mostly accessed by the people, which web pages are accessed together or one after the other or who are accessing what type of web sites and for what purpose. With web usage mining one can find out various patterns such as classifying the users, forming the various clusters based on their similarities, finding association analysis and also can extract personal information about people. The various web logs used in web usage mining are server logs maintained at the
server side by the server, client side web logs maintained by web browsers using cookies or sessions, weblogs maintained by the networks during the transfers by the gateways or proxies etc.

4.2 Architecture of Web usage mining

Architecture of the web usage mining is shown in Fig 1. Initially the data collected from the various sources i.e. server logs is cleaned by data cleaning process to remove the uninteresting data like images, unnecessary links etc is stored into the Clean log. Then interested transactions are identified from the clean log using Transaction Identification process and stores into the Transaction Data. Then it goes through the Integration process by integrating transaction data, user registration data, documentations and usage attributes into integrated data. Then transformation process applies on this data to normalize the data. Various patterns can be extracted from this formatted data like classification, clustering, association analysis etc. Finally Pattern analysis takes place on these patterns using various tools to extract the useful information. The tools used at the final stage are OLAP/Visualization tools to visualize the output in various formats, Knowledge Query Mechanism to extract the required knowledge from the final output by specifying the query, Intelligent agents extracts automatically useful knowledge from the patterns generated.

4.3 Various Sources

Web Usage Mining[1] applications are based on data collected from three main sources[7]: (i) Web servers, (ii) proxy servers, and (iii) Web clients.

4.3.1 The server side

Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. When exploiting log information from Web servers, the major issue is the identification of users_sessions, i.e., how to group all the users_page requests (or click streams) so to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most common approach is to use cookies to track down the sequence of users_page requests for an overview of cookie standards. If cookies are not available, various heuristics can be employed to reliably identify users_sessions. Note however that, even if cookies are used, it is still impossible to identify the exact navigation paths since the use of the back button is not tracked at the server.

Apart from Web logs, users_behavior can also be tracked down on the server side by means of TCP/IP packet sniffers. Even in this case the identification of users_sessions is still an issue, but the use of packet sniffers provides some advantages. In fact: (i) data are collected in real time; (ii) information coming from different Web servers can be easily merged together into a unique log; (iii) the use of special buttons (e.g., the stop button) can be detected so to collect information usually unavailable in log files. Packet sniffers raise scalability issues on Web servers with high traffic, moreover they cannot access encrypted packets like those used in secure commercial transactions. Probably, the best approach for tracking Web usage consists of directly accessing the server application layer. Unfortunately, this is not always possible. First, there are issue related to the copyright of server applications. Most important, following this approach, Web Usage Mining applications
must be tailored for the specific servers and have to take into account the specific tracking requirements.

4.3.2 The proxy side
Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers. Even in this case, session reconstruction is difficult and not all users’ navigation paths can be identified. However, when there is no other caching between the proxy server and the clients, the identification of users’ sessions is easier.

4.3.3 The client side
Usage data can be tracked also on the client side by using Javascript, Java applets, or even modified browsers. These techniques avoid the problems of users’ sessions identification and the problems caused by caching (like the use of the back button). In addition, they provide detailed information about actual user behaviors [15]. However, these approaches rely heavily on the users cooperation and raise many issues concerning the privacy laws, which are quite strict.

4.4. Advantages and Drawbacks of using Web usage mining
There are various benefits and drawbacks of using web usage mining. They are:

4.4.1 Advantages
Web usage mining essentially has many advantages which makes this technology attractive to corporations including the government agencies.

- This technology has enabled e-commerce to do personalized marketing, which eventually results in higher trade volumes.
- The government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of the mining application can benefit the society by identifying criminal activities.
- The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster.
- The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

4.4.2 Draw backs
- Web usage mining by itself does not create issues, but this technology when used on data of personal nature might cause concerns.
- The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. The obtained data will be analyzed, and clustered to form profiles; the data will be made anonymous before clustering so that there are no personal profiles. Thus these applications de-individualize the users by judging them by their mouse clicks. De-individualization, can be defined as a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits.
- Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user’s interests.
- The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site. This trend has increased the amount of data being captured and traded increasing the likeliness of one’s privacy being invaded. The companies which buy the data are obliged make it anonymous and these companies are considered authors of any specific release of mining patterns. They are legally responsible for the contents of the release; any inaccuracies in the release will result in serious lawsuits, but there is no law preventing them from trading the data.
- Some mining algorithms might use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes. This process could result in denial of service or a privilege to an individual based on his race, religion or sexual orientation, right now this situation can be avoided by the high ethical standards maintained by the data mining company. The collected data is being made anonymous so that obtained data and the obtained patterns cannot be traced back to an individual. It might look as if this poses no threat to one’s privacy, actually many extra information can be inferred by the application by combining two separate unscrupulous data from the user.
4.5 Steps in Web Usage Mining

a. Pre-Processing
b. Pattern Discovery
c. Pattern Analysis

4.5.1 Preprocessing
- Data preprocessing has a fundamental role in Web Usage Mining applications.
- The preprocessing of Web logs is usually complex and time demanding.
- It comprises four different tasks:
  (i) Data cleaning
  (ii) Identification and the reconstruction of user sessions
  (iii) Retrieving of information about page content and structure, and
  (iv) Data formatting.

4.5.1.1 Data cleaning
This step consists of removing all the data tracked in Web logs that are useless for mining purposes e.g.: requests for graphical page content (e.g., jpeg and gif images); requests for any other file which might be included into a web page; or even navigation sessions performed by robots and Web spiders. While requests for graphical contents and files are easy to eliminate, robots and Web spiders navigation patterns must be explicitly identified. This is usually done for instance by referring to the remote hostname, 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, an heuristic based on navigational behavior can be used to separate robot sessions from actual user sessions.

4.5.1.2 Session identification and reconstruction
This step consists of
  (i) Identifying the different user sessions from the usually very poor information available in log files
  (ii) Reconstructing the user's navigation path within the identified sessions. The complexity of this step can vary a lot depending on the quality and on the quantity of the information available in the Web logs.

4.5.1.3 Content and structure retrieving
The vast majority of Web Usage Mining applications use the visited URLs as the main source of information for mining purposes. URLs are however a poor source of information since, for instance, they do not convey any information about the actual page content. To employ content based information to enrich the Web log data, introduced an additional categorization step in which Web pages are classified according to their content type; this additional information is then exploited during the mining of Web logs. If an adequate classification is not known in advance, Web Structure Mining techniques can be employed to develop one.

As in search engines, Web pages are classified according to their semantic areas by means of Web Content Mining techniques; this classification information can then be used to enrich information extracted from logs. In Semantic Web for Web Usage Mining: Web pages are mapped onto ontologies to add meaning to the frequently observed paths. Given a page in the Web site, we must be able to extract domain-level structured objects as semantic entities contained in this page. This task may involve the automatic extraction and classification of objects of different types into classes based on the underlying domain ontologies. The domain ontologies themselves may be pre-specified, or may be learned automatically from available training data. Given this capability, the transaction data can be transformed into a representation which incorporates complex semantic entities accessed by users during a visit to the site. Concept-based paths as an alternative to the usual user navigation paths; concept-based path are a high level generalization of usual path in which common concepts are extracted by means of intersection of raw user paths and similarity measures.

4.5.1.4 Data Formatting
This is the final step of preprocessing. Once the previous phases have been completed, data are properly formatted before applying mining techniques stores data extracted from Web logs into a relational database using a click fact schema, so as to provide better support to log querying finalized to frequent pattern mining.

4.5.2 Pattern Discovery
Pattern discovery draws upon the methods and algorithms developed from the several fields such as data mining, pattern recognition, machine learning, and statistics. Methods taken from the other fields taken into consideration for usage of web mining. For example association analysis of Market basket analysis used for the transactions is used here to find the sequential page accesses instead of item accesses. Various patterns used in web mining are statistical analysis, association rules, clustering, classification, sequential patterns and dependency modeling.

4.5.3 Pattern Analysis
Pattern Analysis is to filter out uninteresting rules or patterns from the patterns which discovered in the pattern discovery. The most common form of pattern analysis consists of knowledge query mechanism like SQL.
Another method is using OLAP operations. Visualization can be applied to pattern analysis to present the knowledge effective way.

5. Applications
The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize Web users). This information can be exploited later to improve the Web site from the users’ viewpoint. The results produced by the mining of Web logs can used for various purposes:
To personalize the delivery of Web content;
(i) To improve user navigation through pre-fetching and caching;
(ii) To improve Web design; or in e-commerce sites

5.1 Personalization of web content
Web site personalization[6] can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantage of the user’s navigational behavior. The steps of a Web personalization process include: (a) the collection of Web data, (b) the modeling and categorization of these data (preprocessing phase), (c) the analysis of the collected data, and (d) the determination of the actions that should be performed. The ways that are employed in order to analyze the collected data include content-based filtering, collaborative filtering, rule-based filtering, and Web usage mining.

5.2 Pre-fetching and caching
The results produced by Web Usage Mining can be exploited to improve the performance of Web servers and Web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time.

5.3 Web design
This helps in the designing the web structure based on the users query and shows the more relevant pages to the users instead of more irrelevant to them. This ranks the pages and forms the structure based on the ranks. Usually for any user given query to avoid this web design can help to retrieve only relevant documents.

5.4 E-commerce
Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure. As web usage mining retrieves the user information from the various logs, this information may be related to their personal information, their interests, their economy, their usage time, frequently accessing sites, their buying behavior etc. This information useful in e-commerce for various reasons like, to improve their marketing, to improve their production, to improve customer relationships etc.

Implications of Web Usage Mining for E-commerce
- Determine life time value of clients
- Design cross marketing strategies across products
- Evaluate promotional campaigns
- Predict user behavior based on user’s profile and previous rules
- Present dynamic information to users based on their interests and profiles
- Present ads and coupons to users based on their access patterns
- Proposing promotions based on their regular buying behavior

6. Tools used for Web Usage Mining
Various tools used for web usage mining are Web Utilization Minor (WUM), Web Site Information Filter System (WebSIFT)[5], KOINOTITES used for Web personalization etc.

6.1 Web Utilization Miner(WUM)
Web Utilization Miner WUM, a mining system for the discovery of interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the human expert using WUM’s mining language MINT. MINT supports the specification of criteria of statistical, structural and textual nature. To discover the navigation patterns satisfying the expert’s criteria, WUM exploits an innovative aggregated storage representation for the information in the web server log.
Knowledge about the navigation patterns occurring in or dominating the usage of a web site can greatly help the site’s owner or administrator in improving its quality. Data mining can assist in this task by effectively extracting knowledge from the past, i.e. from the site access recordings.

The web miner proposed in [2] simply discovers statistically dominant paths. The “WEBMINER” tool of [3] provides a query language on top of external mining software for association rules and for sequential patterns. Web Utilization Miner WUM employs an innovative technique for the discovery of navigation patterns over an aggregated materialized view of the web log. This technique offers a mining language as interface to the expert, so that the generic characteristics can be given, which make a pattern interesting to the specific person. Thus, only patterns having the desired characteristics are constructed, while uninteresting patterns are pruned out early.

There are two major modules: the Aggregation Service prepares the web log data for mining and the MINT-Processor does the mining.

The Aggregation Service extracts information on the activities of the users visiting the web site and groups consecutive activities of the same user into a transaction. It then transforms transactions into sequences.

The MINT-Processor mines the aggregated data according to the directives of the human expert.

“MINT” is the mining language serving as interface between the user and the miner. The expert uses MINT to instruct the miner on the formulation of the output, and, most importantly, on the interestingness criteria to be satisfied by the desired patterns.

6.2 The Web Site Information Filter System (Web SIFT)

Web usage mining is the application of the data mining technique to large web data repositories in order to extract the web usage patterns. For Web usage mining, there are three types of domain information available: usage, content and structure. Web SIFT system uses content and structure information from the web site in order to identify potentially interesting results from the mining usage data. This is a framework for web usage mining to perform preprocessing and knowledge discovery and automatically defines a belief set. The information filter uses this belief set to find interesting patterns.

Web SIFT system is based on the Web Miner prototype [3] divides the web usage mining into three parts: Preprocessing, Pattern Discovery and Pattern Analysis. Web SIFT system implemented using a Relational database, Procedural SQL and Java programming language.

6.3. KOINOTITES: A Web Usage Mining Tool for Personalization

This tool uses data mining techniques to construct user communities on the web. User communities model groups of visitors in a Web site, who have similar interests and navigational behavior. We present the architecture of the system and the results that we obtained in a real Web site. The personalization of Web services is a leap in the direction of alleviating the information overload problem and making the Web a friendlier environment for its users. As stated in [4]: "...the Web is ultimately a personal medium in which every user's experience is different than any other's". Web Personalization [5] is the task of making Web based information systems adaptive to the needs and interests of individual users, or groups of users. Typically, a personalized Web site recognizes its users, collects information about their preferences and adapts its services, in order to match the users’ needs. KOINOTITES, is a software system that exploits Web Usage Mining and user modeling techniques for the customization of information to the needs of individual users. More specifically, KOINOTITES processes the Web server log files, and organizes the information of a Web site (i.e., Web pages), into groups, which reflect common navigational behavior of the Web site visitors.

KOINOTITES is a software tool, which exploits Web Usage Mining techniques in order to create user communities from Web data. KOINOTITES is based on a modular architecture, and comprises the following two main components:

i) A mining component, that consists of the modules that perform the main functions of the system, i.e., data preprocessing, session identification, pattern recognition and knowledge presentation.

ii) A Graphical User Interface (GUI) component, supplemented by wizards and on-line help that is used for user interaction with the system.

Both components have been implemented using the Java programming language.

7. CONCLUSIONS

This paper presented an overview of the Web usage mining to extract the usage patterns of the users over the web from the various log file. This also specified various data mining techniques used in the web usage mining, applications of the web usage mining and various tools used in web mining. In conclusion, we believe that Web Usage Mining is a very promising solution that can help in producing personalized Web-based systems, making access to on-line information more efficient. This issue is becoming crucial as the size of the Web increases at
breathtaking rates.

References
The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage: http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digital Library, NewJour, Google Scholar