WEB MINING IN E-COMMERCE

Istrate Mihai
University of Pitești Faculty of Mathematics and Informatics No.1, Stadionului, Țicleni, Gorj, 215600, Romania mihaifrance@yahoo.com +40 745 775 935

Recently, the web is becoming an important part of people's life. The web is a very good place to run successful businesses. Selling products or services online plays an important role in the success of businesses that have a physical presence, like a retail business. Therefore, it is important to have a successful website to serve as a sales and marketing tool. One of the effective used technologies for that purpose is data mining. Data mining is the process of extracting interesting patterns from large databases. Web mining is the usage of data mining techniques to extract interesting information from web data. This paper presents the three components of web mining: web usage mining, web structure mining and web content mining and the main data preprocessing tasks for web usage mining.

Keywords: E-Commerce, Data mining, Web mining
JEL M

1. E-Commerce and Retail websites

In e-commerce instead of having your business in a limited physical place and a limited sector of customers who are usually near to your store or business, you have it in the web. In e-commerce websites you have the ability to sell, advertise, and introduce different kinds of services and products in the web. E-commerce websites have the advantage of reaching a large number of customers regardless of distance and time limitations. Furthermore, an advantage of e-commerce over traditional businesses is the faster speed and the lower expenses for both e-commerce website owners and customers in completing customers transactions and orders.

Because of the above advantages of e-commerce over traditional businesses, a lot of industries in different fields such as retailing, banking, medical services, transportation, communication, and education are establishing their business in the web. But creating a successful online business can be a very difficult and costly task if not taking into account e-commerce website design principles, web engineering techniques, and what e-commerce is supposed to do for the online business. Understanding the requirements of both e-commerce website owner and customer is an important aspect in building a successful e-commerce website. There is a lot of information need to be defined before starting building the e-commerce website such as identifying business goals and how the website will target those goals, if the website supposed to attract new customers or increase the sales of current customers, identify if the proposed website will increase the business overall profit, and identify the most suitable tools and techniques need to be used/followed in order to target those requirements.

Retail websites aim to inspire, reflect a good image about the business and improve it online. An important factor in having a successful retail website is to know your competitors. On one hand, by identifying their points of strongness and trying to get benefit of them by improving those strongness points and adopting powerful strategies. On the other hand, identifying weakness points of your competitors and avoid them is a good practice in having a successful retail website.

2.Web mining

The usage of data mining to maintain websites and improve their functionality is an important field of study. Patterns extracted from applying data mining techniques on web data can be used to maintain websites by improving their usability through simplifying user navigation and information accessibility and improving the content and the structure of the website in a way that meets the requirements of both website owner and user which will consequently increase the overall profit of the business.

Web mining is the use of data mining techniques to extract useful patterns from the web. Those extracted patterns are used to improve the structure of websites, improve the availability of the information in the websites and the way those

pieces of information are introduced to the website user, and to improve data retrieval and the quality of automatic search of information resources available in the web. Web mining can be divided into three major categories: web usage mining, web content mining, and web structure mining.

2.1 Web Usage Mining

Web usage mining or web log mining is the process of applying data mining techniques to web log data in order to extract useful information from user access patterns. Web usage mining tries to make sense of the data generated by the web user's sessions or behaviors. The web usage data includes data from web server access log, proxy server logs, browser logs, user profiles, registration data, cookies, and user queries. Web usage mining tries to predict user behavior while user interacts with the web and learns user navigation patterns. The learned knowledge could then be used for different applications such as website personalization, business intelligence, usage characterization and adaptive websites. There are two approaches for web usage mining process:

- Mapping the log data into relational tables before an adopted data mining techniques is performed.
- Using the log data directly by utilizing special preprocessing techniques.

The Web usage mining process consists of three phases: data preprocessing, pattern discovery, and pattern analysis. Pattern discovery is that set of methods, algorithms, and techniques used to extract patterns from web log file. Several
techniques are used for pattern discovery such as statistical analysis, clustering, classification, and sequential pattern mining. After patterns are discovered they need to be analyzed in order to determine interesting and important patterns, besides the removal of redundant patterns. Pattern analysis has several different forms such as knowledge query mechanism, visualization techniques, and loading usage data into a data cube in order to perform Online Analytical Processing OLAP operations.

A web server log file records users transactions in the web. Usually, the web log file contains information about the user IP address, the requested page, time of request, the volume of the requested page, its referrer, and other useful information. The web log file can have different format, but there is a common log file format that is mostly used. The common log file has the following format:

remotehost rfc931 authuser [date] "request" status bytes

where *remotehost* represents remote hostname (or IP number if DNS hostname is not available), *rfc931* represents the remote logname of the user, *authuser* represents the username as which the user has authenticated himself, [*date*] represents dateand time of the request, *"request"* represents the request line exactly as it came from the client, *status* represents the HTTP status code returned to the client, and finally *bytes* represents the content-length of the document transferred. The WWW Consortium (W3C) presented an extended format for web server log file that is able to record a wide range of data to make an advanced analysis of the web log file. Web log file is the main source of data analysis in web mining but a lot of preprocessing efforts need to be performed in order to prepare the web log file to be mined.

2.2 Web Content Mining

Web content mining is mining the data that a web page contains. The contents of most of the web pages are texts, graphics, tables, data blocks, and data records. A lot of research has been done to cover different web content mining issues for the purpose of improving the contents of the web pages, improving the way they are introduced to the website user, improving the quality of search results, and extracting interesting web page contents.

Web content mining is still a large field. It constains:

- structured data extraction;
- sentiment classification, analysis and summarization of consumer reviews;
- information integration and schema matching;
- knowledge synthesis;
- template detection and page segmentation;

A large amount of information on the Web is contained in regularly structured data objects which are data records retrieved from databases. Such Web data records are important because they often present the essential information of their host pages, lists of products and services.

Two of the most used methods for extracting structured data are *wrapper induction* (given a set of manually labeled pages, a machine learning method is applied to learn extraction rules or patterns) and *automatic extraction* (given a set of positive pages or given only a single page with multiple data records, generate extraction patterns).

2.3 Web Structure Mining

Links pointing to a document indicate the popularity of the document, whereas links coming out of a document indicate the richness or the variety of topics covered in the document. Web structure mining describes the organization of the content of the web where structure is defined by ”hyperlinks between pages and HTML formatting commands within a page”.

Understanding the relationship between contents and the structure of the website is useful to keep an overview about websites. One of the approach allows the comparison of web page contents with the information implicitly defined by the structure of the website. In this way, it can be indicated whether a page fits in the content of its link structure, and identify topics which span over several connected web pages. Thus supporting web designers by comparing their intentions with the actual structure and content of the web page. Other studies deal with the web page as a collection of blocks or segments. By partition the web page into blocks and by extracting the page-to-block, block-to-page relationship from link structure and page layout analysis, a semantic graph can be constructed over the WWW such that each node exactly represents a single semantic topic, this graph can better describe the semantic structure of the web. Structure within a web page can be used to help machines understand pages.

3.Web Usage Mining Techniques

In this section, we discuss data mining techniques that are mostly used in web usage mining such as statistical analysis techniques, clustering, classification, association rule mining, and sequential pattern mining.

Statistical analysis is the process of applying statistical techniques on web log file to describe sessions, and user navigation such as viewing the time and length of a navigational path. Statistical prediction can also be used to predict when some page or document would be accessed from now. It makes use of the N-grammer model which assumes that when a user is browsing a given page, the last N pages browsed affect the probability of the next page to be visited.

Clustering is the process of partitioning a given population of events or items into sets of similar elements. In web usage mining there are two main interesting clusters to be discovered: usage clusters, and pages clusters. An approach is to cluster web pages to have a high quality clusters of web pages and use that clusters to produce index pages, where index pages are web pages that have direct links to pages that may be of interest of some group of website navigators.

Classification is dividing an existing set of events or transactions into another predefined sets or classes based on some characteristics. In web usage mining, classification is used to group users into predefined groups with respect to their

navigation patterns in order to develop profiles of users belonging to a particularclass or category.

Association rule mining is the discovery of attribute values that occur frequently together in a given set of data. Association rules mining techniques are used in web usage mining to find pages that are often viewed together, or to show which pages tend to be visited within the same user session. A re-ranking method with the help of website taxonomy is to mine for generalized association rules and abstract access patterns of different levels to improve the performance of site search. Another approach for predicting web log accesses is based on association rule mining. Association rule mining facilitates the identification of related pages or navigation patterns which can be used in web personalization.

In sequential pattern mining a sequence of actions or events is determined with respect to time or other sequences. In web usage mining, sequential pattern mining could be used to predict user's future visit behaviors. Some web usage

mining and analysis tools use sequential pattern mining to extract interesting patterns such as SpeedTracer and Webminer.

4. Data Preprocessing for Web Usage Mining

Before data mining techniques are applied to web log file data, several preprocessing steps should be done in order to make web log file data ready to be mined. Web log file contains data about requested URL, time and date of request, method used, etc. The main data preprocessing tasks are data cleaning and *filtering, path completion, user identification, session identification,* and *session formatting.*

Data cleaning is the first preprocessing task. It involves the removal or elimination of irrelevant items that are not important for any type of web log analysis.Elimination of irrelevant items can be accomplished by checking the suffix of the URL name to filter out requests for graphics, sound, and video hits in order to concentrate on data representing actual page hits. For example, all log entries with filename suffixes such as gif, jpeg, and jpg can be removed. Another cleaning process is removing log entries generated by web agents like web spiders, indexers, or link checkers. Filtering out failed server requests, or transforming server error code is also done. Merging logs from multiple servers and parsing the log into data fields is also considered a data cleaning step.

Path completion preprocessing task fills in page references that are missing due to local browsing caching such as using the back button available in the browser to go back to previously visited page.

User identification is a complex step due to the existence of local caches, corporate firewalls, and proxy servers. If the agent log shows a change in browser software, or operating system, a reasonable assumption to make is that each different IP address in the log file represent a different user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, a heuristic assume that there is another user with the same IP address. Another assumption can be made is that consecutive accesses from the same host during a certain time interval come from the same user. In some cases it is difficult to identify users, for example, when two users use the same machine and the same browser with the same IP address and look at the same set of pages.

Session identification. A user session is defined as "the set of pages visited by the same user within the duration of one particular visit to a website". Session identification is dividing the page accesses of each user into individual sessions. One approach to identify user sessions, is by using a timeout threshold that is if the time between pages requests exceeds a certain limit, then the user is starting a new session. Another approach assumes that consecutive accesses within the same time period belong to the same session.

Session Formatting. A final preprocessing step could be formatting the sessions or transactions for the type of the data mining technique, or algorithm to be applied. The Webminer, for example, formats the cleaned web server log data in order to apply either association rule mining or sequential pattern mining.

5. Discussion

From previous, it is clear that making changes and adaptations to websites with the help of extracted patterns using different data mining techniques is very effective, but doing that in the maintenance phase can be costly and time consuming and suffers from different drawbacks. In commercial companies which are companies that sell different kinds of products on the web, in order to make an effective maintenance to their websites, the companies have to wait some period of time, for example one year, in order to have a representative log file that reflects customers transactions in their website and can give a clear image about their behavior. This amount of time is considered very big especially for the companies in which the time factor plays an important role in their success strategy, and have many competitors who can attract their customers if they have no solid marketing strategies in order to keep their customers as loyal as possible.

On the other hand, most businesses gather information about internet customers through online questionnaires. But, many customers choose not to complete these questionnaires because of the amount of time required to complete them as well as a lack of a clear motivation to complete them. Several companies use cookies to follow customers through the WWW, but cookies are sometimes detected and disabled by web browsers and do not provide much insight into customer preferences. This is because customers are feeling that their profiles are not secure so a number of customers choose to give incorrect information about themselfs.

Furthermore, in web mining different strategies are implemented to identify sessions such as defining a time threshold that a session should not exceed or assuming that consecutive accesses within the same time period belong to the same session. In some cases, it is difficult to identify users, for example, when two users use the same machine and the same browser with the same IP address and look at the same set of pages. We can conclude from that, that those session and user identification strategies can not give a guarantee that those identified users and sessions represent the actual users and sessions.

The problem of building an ill-structured website for some company/business can be solved by applying data mining techniques such as clustering, classification, and association rule mining on the contents of the information system of the company/business. Then, from the extracted patterns, the information needs to be considered in the website building process is gained and invested during the design phase in the process of website design which yields to a better designed retail website. The main advantage of this method is that it reduces maintenance time and budgetary costs for websites if they are built taking into account the extracted interesting patterns from the transactions database of the company/business. This approach also permits the sales manager to focus on the core business and gives him a better view about his products and customers which is very helpful in designing retail websites.

In conclusion, patterns extracted from applying web mining techniques on web data can be used to maintain websites by improving their usability through simplifying user navigation and information accessibility and improving the content and the structure of the website in a way that meets the requirements of both website owner and user which will consequently increase the overall profit of the business.

References

1. S. Ananyan, M. Kiselev, Automated Analysis of Unstructured Texts
2. Michael Goebel, Le Gruenwald, A Survey of Data Mining and Knowledge
3. Stefan Conrad, Martin Mauve, Data Mining for Retail Website Design and Enhanced Marketing
4. Bing Liu, Web Content Mining
5. Christopher J. Hazard, Data Mining and Web Logs
6. Asem Omari and Stefan Conrad, Web Usage Mining for Adaptive and Personalized Websites
7. Margaret H. Dunham, Data Mining Introductory and Advanved Topics
8. Ruey-Shun Chen, Ruey-Chyi Wu, and J. Y. Chen, Data Mining Application in Customer Relationship Management of Credit Card Business.
9. A. McDonald and R. Welland, Web Engineering in Practice
10. Martin Ester, Hans-Peter Kriegel, and Matthias Schubert, Web Site Mining a New Way To Spot Competitors, Customers and Suppliers in The World Wide Web.
11. M. Istrate, Web Content Mining.