Research on Web Usage Mining for Electronic Commerce

Li Chaofeng\(^1\), Lu Yansheng\(^1\)

\(^1\) College of Computer, Huazhong University of Science and Technology, Wuhan 430074, P.R. China
\(^2\) Colleges of Management, South-Central University for Nationalities, Wuhan 430074, P.R. China

Abstract Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. On the other hand, the rapid growth of e-commerce has caused product overload where customers on the Web are no longer able to effectively choose the products they are exposed to. Applying Web usage mining techniques to e-commerce will be a solution to overcome these problems. This paper has attempted to provide a survey of the rapidly growing area of Web usage mining. We propose a framework of Web usage mining for e-commerce, discuss and analyze each module of the framework in detail, give some applications of which to e-commerce. Finally, we conclude this paper and propose the future research directions.

Keywords Web Usage Mining; Electronic Commerce; Data Preprocessing; Patterns discovery

1 Introduction

E-commerce has been growing rapidly keeping the pace with the Web. At the same time it is generating huge amount of data, e.g. customer purchases, browsing patterns, usage times and preferences, at an increasing rate. Moreover, its rapid growth has made both enterprise and customers face a new situation: enterprise is difficult to survive due to more and more competitions. On the other hand, for customers, before they determine which products meet their needs, the opportunity to choose among more and more products has increased the burden of information processing. As a result, it is necessary for us to make new marketing strategies such as one-to-one marketing and customer relationship management (CRM). One solution to realize these strategies is personalized recommendation that helps customers find the products they would like to purchase by producing a list of recommended products for each given customer.

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining has been an important technology for understanding users' behaviors on the Web. It can assist e-commerce to seek and retain the most profitable customers by analyzing demographic data, customer buying and traversing patterns collected online or offline. In addition, e-commerce companies can improve products quality or sales by anticipating problems before they occur with the use of Web usage mining techniques. They also provide companies with previously unknown buying patterns and behavior of their online customers. More importantly, the fast feedback the companies obtain by using Web usage mining is very helpful in increasing the company's benefit.

The research on Web usage mining associate with e-commerce is a focus field nowadays. This paper attempts to present Web usage mining processes and issues in e-commerce applications. The organization of our paper is as following. In section 2, we propose a framework of Web usage mining, discuss and analyze the main techniques of each step in detail. Some applications of Web usage mining techniques to e-commerce are given in section 3. Finally, section 4 concludes this paper with suggestions for future research.

2 Web Usage Mining

Web usage mining is the process of applying data mining techniques to the discovery of behavior patterns based on Web data, for various applications. The goal of Web usage mining is to capture and model the behavioral patterns and profiles of users interacting with a Web site. In the advance of
e-commerce, the importance of Web usage mining grows larger than before.

2.1 Web Data and Data Collection of Web Usage Mining

There are many kinds of data that can be used in Web usage mining. Each type of data collection differs in the location of the data source and the kinds of data available, the segment of population from which the data was collected, as well as its method of implementation. According to Srivastava et al.\[^8\], these data can be classified in four categories:

1. Content. Data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.

2. Structure. Data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.

3. Usage. Data represent a Web site’s usage, such as a visitor’s IP address, time and date of access, complete path accessed, and other attributes that can be included in a Web access log.

4. User profile. Data provide information about the users of a Web site. A user profile contains demographic information such as name, age, country, marital status, education, interests etc.

Web usage mining’s data source can be collected at the server-side, client-side, proxy servers, or obtained from an organization’s database which contains business data or consolidated Web data.

2.1.1 Server Level Collection

Since a Web server log explicitly records the browsing behavior of site visitors, it is an important data source for performing Web usage mining. The data recorded in server logs reflects the access of a Web site by multiple users. However, the site usage data recorded by server logs may not be entirely reliable due to many reasons, such as the cached page views, use of POST method to transfer information and use of cookies having raised growing concerns regarding user privacy etc. Otherwise, the Web server also relies on other utilities such as CGI scripts to handle data sent back from client browsers.

2.1.2 Client Level Collection

Client-side data collection can be implemented by two ways: using a remote agent such as Java scripts or Java applets and modifying the source code of an existing browser such as Mosaic or Mozilla to enhance its data collection capabilities. Client-side collection has an advantage over server-side collection because it improves both the caching and session identification problems. However, the implementation of which requires user cooperation. Java applets may generate some additional overhead especially when they are loaded for the first time. Java scripts cannot capture all user clicks (such as reload or back buttons) either. The method by modifying the source code of an existing browser must convince the users to use the modified browser for their daily browsing activities, but it is very difficult.

2.1.3 Proxy Level Collection

A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides\[^9\]. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

2.2 A Framework for Web Usage Mining

The primary objective of Web mining is to discover interesting patterns in accesses to various Web pages within the Web space associated with a particular server. Generally, Web usage mining comprises three phases, namely preprocessing, pattern discovery, and pattern analysis, as shown in figure 1.

2.2.1 Data Preprocessing

As discussed in section 2.1, the data collected are different in both locations and types. Otherwise, lack of structure, heterogeneity and volumes are the main characteristics of the Web data. So, it is necessary to preprocess the data collected before applying Web usage mining tools to discovery patterns.
Generally, data preprocessing consists of data cleaning, user identification, session identification, path completion and formatting.

(1) Data Cleaning. This technique can eliminate irrelevant and redundant data. It is very important for any type of pattern discovery’s tools. Elimination of the items deemed irrelevant can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG, and map can be removed. In addition, common scripts such as the files requested with the suffixes of “.cgi” can also be removed\(^ {10}\). But, in some cases, there are important accesses that are not recorded in the access log, such as local caches and proxy servers can severely distort the overall picture of user traversals through a Web site. Current methods to overcome this problem include the use of cookies, cache busting, and explicit user registration.

(2) User Identification. A user is defined as the principal using a client to interactively retrieve and render resources or resource manifestations. User identification is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The Web usage mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, even for the existence of local caches, corporate firewalls, and proxy servers. The Web usage mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, even for the log/site-based methods, there are heuristics that can be used to help identify unique users. For example, even if the IP address is the same, if the agent log shows a change in browser software or operating system, a reasonable assumption to make is that each different agent type for an IP address represents a different user\(^ {11}\). Another heuristic for user identification is to use the access log in conjunction with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, the heuristic assumes that there is another user with the same IP address.

(3) Session Identification. A user session means a delimited set of user clicks (click stream) across one or more Web servers. The goal of session identification is to divide the page accesses of each user

![Figure 1 A framework for Web usage mining](image)
into individual sessions. The simplest method of achieving this is through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout. According to the empirical data having done by Catledge L and Pitkow J., a timeout of 25.5 minutes may be better than 30 minutes. Another method is using the concept of maximal forward reference to identify user session.

(4) Path Completion. Path completion resolves the problem that there are important accesses that are not recorded in the access log. Methods similar to those used for user identification can be used for path completion. If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. If the page is in the user’s recent request history, the assumption is that the user backtracked with the “back” button available on most browsers, calling up cached versions of the pages until a new page was requested. If the referrer log is not clear, the site topology can be used to the same effect. If more than one page in the user’s history contains a link to the requested page, it is assumed that the page closest to the previously requested page is the source of the new request.

(5) Formatting. Once the appropriate preprocessing steps have been applied to the server log, a final preparation module can be used to properly format the sessions or transactions for the type of data mining to be accomplished. For example, since temporal information is not needed for the mining of association rules, a final association rule preparation module would strip out the time for each reference, and do any other formatting of the data necessary for the specific data mining algorithm to be used.

2.2.2 Pattern Discovery

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. There are many different kinds of pattern discovery tools can be used when dealing with Web usage mining. Here we analyze some of them as following:

(1) Statistical Analysis. Statistical techniques are the most common method to extract knowledge about visitors to a Web site. Even without the benefit of an integrated e-commerce datamart, statistical analysis can be performed on the preprocessed session or transaction data. Many Web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This report may include limited low-level error analysis such as detecting unauthorized entry points or finding the most common invalid URI. Despite lacking in the depth of its analysis, this type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions.

(2) Association Rules. Association rules capture the relationships among items based on their patterns of cooccurrence across transactions (without considering the ordering of items). In the case of Web transactions, association rules capture relationships among pageviews based on the navigational patterns of users. Most common approaches to association discovery are based on the Apriori algorithm that follows a generate-and-test methodology. Aside from being applicable for business and marketing applications, the presence or absence of association rules can help Web designers to restructure their Web site. The association rules may also serve as a heuristic for prefetching documents in order to reduce user-perceived latency when loading a page from a remote site.

(3) Clustering. Clustering is a technique to group together a set of items having similar characteristics. In the Web usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-commerce applications or provide personalized Web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers.

(4) Classification. Classification is the task of mapping a data item into one of several predefined classes. For this task, a model set (i.e., a set of cases whose class labels are known) is first analyzed.
and a classification model is constructed based on the features available in the data of the model set. Such a classification model is then used to categorize a score set (i.e., a set of cases whose class labels are unknown). Classification can be done by using supervised inductive learning algorithms such as decision tree, naive Bayesian, k-nearest neighbor classifiers, Support Vector Machines etc.

(5) Sequential Patterns. Sequential patterns in Web usage data capture the Web page trails that are often visited by users, in the order that they were visited. Sequential patterns are those sequences of items that frequently occur in a sufficiently large proportion of transactions. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. It may be useful in trend analysis, change point detection, or similarity analysis.

(6) Dependency Modeling. The goal of dependency modeling is to develop a model that can represent significant dependencies among the various variables in the Web domain. For example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (i.e. from a casual visitor to a serious potential buyer). Hidden Markov Models and Bayesian Belief Networks are the methods that often employed to model the browsing behavior of users. Modeling of Web usage patterns will provide a theoretical framework for analyzing the behavior of users, as well as is potentially useful for predicting future Web resource consumption. Such information may help develop strategies to increase the sales of products offered by the Web site or improve the navigational convenience of users.

2.2.3 Pattern Analysis

The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop techniques and tools for helping an analyst better assimilate it. Pattern analysis tools can do this. The most common tools of pattern analysis include visualization techniques, OLAP analysis and knowledge query mechanism such as SQL-like mechanism.

(1) Visualization. Visualization is a natural choice for understanding the behavior of Web users, for it has been used very successfully in helping people understand various kinds of phenomena, both real and abstract. As an example, WebViz system is used to visualize WWW access patterns[17]. It allows the analyst to selectively analyze the portion of the Web that is of interest by filtering out the irrelevant portions.

(2) OLAP. During the research of pattern analysis, the researcher found that the analysis needs of Web usage data have much in common with those of a data warehouse, and hence OLAP techniques are quite applicable[18]. The implementation of this method is to load usage data into a data cube in order to perform OLAP operations.

(3) Knowledge Query Mechanism. Knowledge query mechanism mainly implemented association with SQL. Many Web usage analysis tools, e.g. WUM, WebMiner and Midas, give some object rules, e.g. support and confidence that are helpful to filter out unimportant knowledge manually, and then acquire the analyzed results by using SQL-like.

3 Applications to E-commerce

As discussed before sections, the results of Web usage mining can be applied to understand and analyze Web usage data. So, we can apply web usage mining techniques to e-service, especially e-commerce. The following are some examples.

3.1 Personalization

The tremendous growth in the number and the complexity of information resources and services on the e-commerce site has made Web personalization an indispensable tool for both Web-based organizations and for the end users. The ability of an e-commerce site to engage visitors at a deeper level, and to successfully guide them to selection and purchase products is now viewed as one of the key factors for e-commerce company ultimate success. Making dynamic recommendations to a Web user, based on her/his profile in addition to usage behavior is very attractive to many applications, e.g.
cross-sales and up-sales in e-commerce. Web usage mining is an excellent approach for achieving this goal. Currently, there are many systems have already concentrated on providing Web Site personalization based on usage information, such as The WebWatcher [19], SiteHelper [20] and Letizia [21] etc. The main advantage of using Web personalization based on Web usage mining is that it can automate the adaptation of Web-based services to their users. This overcomes the problem in traditional Web personalization systems, i.e. the personalization process involves substantial manual work and most of the time, significant effort on the part of the user.

3.2 System Improvement

Every customer using e-commerce system expects that the system has high performance and other service quality attributes such as databases, networks, etc. Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission, load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as e-commerce continues to grow at an exponential rate. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins etc.

3.3 Site Modification

The attractiveness of a Web site, in terms of both content and structure, is very important for a product catalog of e-commerce. Web usage mining provides detailed feedback on user behavior, providing the Web site designer information on which to base redesign decisions. The Web site designer can use this information to change the structure of a site to adapt to the customers using his/her e-commerce system. Clustering of pages is used to determine which pages should be directly linked.

3.4 Business Intelligence

Information on how customers are using a Web site is critical information for marketers of e-commerce. By mining the relationship between customers’ behavior and purchase, we can understand the customers’ purchasing intention much better, find the customers’ purchasing characteristics and trends, and identify the potential purchaser. According to these, we can perform business intelligence, support business intelligence and determine advertisement policy reasonably. There are several commercial products, such as SurfAid, Accrue, Net-Genesis, Aria, Hitlist and WebTrends that provide Web traffic analysis mainly for the purpose of gathering business intelligence.

4 Conclusion and Future Research Directions

This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web usage mining. We proposed a framework of Web usage mining. Next, we discussed the process of Web usage mining according to our framework in detail. Finally, we give some examples that applying Web usage mining to e-commerce. With the growth of Web-based applications, specifically e-commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. However, Web Usage mining raises some hard scientific questions that must be answered before robust tools can be developed. The following are the problems that we should research in the future:

(1) Data Preprocessing. There are three questions must be consideration: first, there will be a continual need to develop better instrumentation and data collection techniques to overcome the conflict between the analysis needs of the analysts and the privacy needs of users. Second, we should examine more intelligent data integration techniques. For portions of Web usage data exist in different sources. Finally, we believe there is a need to group individual data collection events into groups. The future research should pay more attention to this issue.

(2) Patterns Discovery. The key component of Web mining is the mining process itself. A lot of work still remains to be done in adapting known mining techniques as well as developing new ones. Specifically, the following issues must be addressed: new types of knowledge, improved mining
algorithms, incremental Web mining and distributed Web mining.

(3) Patterns Analysis. There is a need to develop tools which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge. In addition, we should develop new intelligent tools that can assist in the interpretation of mined knowledge.

References