Web Mining in E-Commerce: Pattern Discovery, Issues and Applications

Ketul B. Patel[1], Jignesh A. Chauhan[2], Jigar D. Patel[3]

Acharya Motibhai Patel Institute of Computer Studies
Ganpat University, Kherva, India

Abstract— The World Wide Web is the key source of information and it is growing rapidly. E-commerce has provided a cost efficient and effective way of ding business. Web mining is the application of data mining technique to discover useful information from World Wide Web. Web mining is applied to e-commerce to know the browsing behaviour of customers, to determine the success of marketing efforts, to improve the design of e-commerce web site and to provide personalized services. This paper discusses web mining in e-commerce, the categories of web mining, pattern discovery techniques to find out interesting patterns, issues of web mining in e-commerce and application of web mining in e-commerce.

Keywords— E-Commerce, Pattern Analysis, Pattern Discovery, Web Mining.

I. INTRODUCTION

In e-commerce websites, you can 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. The 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 customer 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.

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web [4]. Based on the different emphasis and different ways to obtain information, web mining can be divided into three major categories. The first is web content mining. The knowledge is taken from Web page contents i.e. from the topics of different sites the useful data can be extracted. It is the automatic search and retrieval of information and resources available from millions of sites. The second sub-category is web structure mining. Here the knowledge is taken from hyperlinks and it shows how pages are connected one with another. The third sub-category is web usage mining. It helps to define the behaviour of visitors and classify them into groups.

With the explosion of E-commerce, the way companies are doing businesses has been changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us a cost-efficient and effective way of doing business. Unfortunately, to most companies, web is nothing more than a place where transactions take place. All the e-commerce sites have high traffic. People surf the sites very often but the income is not always very high. So, the web data mining appeared and also nowadays much attention is paid to it. It is very important to apply web data mining to e-commerce in order to gather knowledge about users and rank data accordingly. Web data mining is a branch of data mining. It is advance successful technology through which information is filtered easier. So, web data mining became a publicly accessible source that gives promising results. With the use of e-commerce through internet, companies find a new and better way to do business. After developing the web site thought companies get benefits, they should not sit relaxed. Companies have to implement Web mining systems to understand their customers' profiles and to identify their own strength and weakness of their E-marketing efforts on the web through continuous improvements. Internet is a gold mine, but only for those companies who realize the importance of Web mining and adopt a Web mining strategy now. Web mining technology has many important roles that should be mentioned. It can automatically find, extract information from the variety web resources. It also develops, improves and enhances the quality and the efficiency of search engines, determines web pages or files, makes classifications [7]. It can also generate large-scale real-time data. Web data mining discovers useful information from the Web hyperlink and page content. It has already changed the face of many business functions in a modern competitive enterprise. It is obviously easier to make right business decisions or understand the information that came from customers with the help of web data mining. It helps e-commerce to understand how to improve its services for special groups of customers and clients, and what tasks to realize. The e-commerce site can increase the exposure of its product pages and so average order size can be increased. Companies can save percentage of its budget per month owing to knowledge that was received from web mining analysis. Web data mining gathers implicit knowledge about clients and instructs e-commerce in every aspect. Then, it extracts valuable and comprehensible information from huge web resources to instruct e-commerce. It also gathers the information in an automated way and builds models used to predict customer purchasing decisions. Web mining is very precious to the company in the fields of understanding customer behaviour, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on.

The pattern discovery techniques include algorithms to find interesting and useful patterns from web data. Some of them are association rules, clustering, sequential patterns, classification etc. Pattern analysis techniques are used to highlight overall patterns in data and to filter out uninteresting patterns. The statistical data should be analysed on the types of visitors that come to the website. Also, it is important to analyse the steps visitors make to reach the site. Sometimes, these are similar steps, similar key-words or similar tags. Further, association rules are derived in order to identify the correlations between the web pages. Then it should be analysed how the goal of the client's visit was satisfied. With the vast amount of customers, it is well understood in the web data mining. There are several methods of web data mining in e-commerce. The path analysis method is used to improve the page and site structure. Discovery of association rules explains how the actions are related. This method helps to illuminate unneeded information from web pages. Discovery of sequence models method helps to personalize services through the behaviour of the customers. Discovery of classification rule method explains how to classify users. The discovery of cluster analysis shows how the users are similar to provide better services in e-commerce [8].

II. Web Mining

Web Mining is the application of data mining technique to discover and retrieve useful and interesting patterns from web data [4]. Web data contains different kinds of information including web documents data, web structure data and web log data. According to the kinds of data to be mined, Web Mining can be broadly divided into three categories: Web content mining, Web structure mining and Web usage mining.

A. *Web Content Mining*

Web Content Mining is the task of extracting knowledge from the content of documents on World Wide Web. The Web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of Web data forces the Web content mining technique towards a more complicated approach. Web content mining describes the automatic search of information resources available online, and involves mining Web data content. It focuses on techniques for searching the web for documents whose content meets a certain criterion. The documents found are used to build a local knowledgebase.

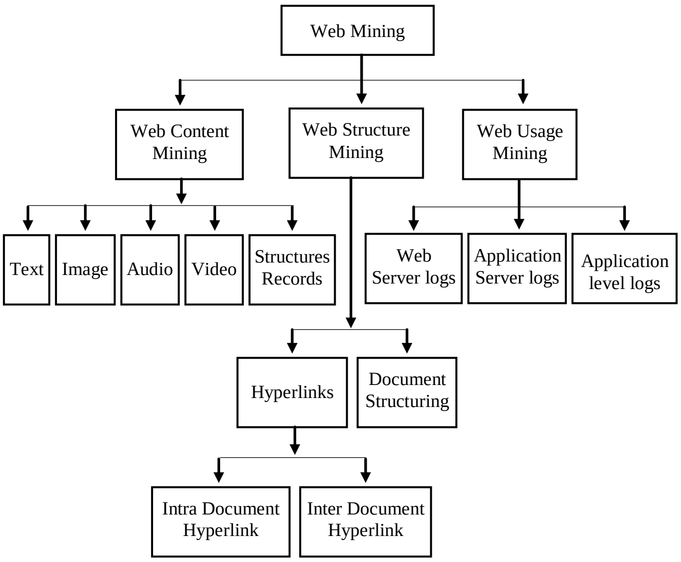

Fig. 1 Web Mining Categories

B. *Web Structure Mining*

Web Structure Mining is the process of discovering structure information from the Web. It describes the connectivity in the Web subset based on the given collection of interconnected Web documents. The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink. Mining the structure and Web page structure, can be used to guide the classification and clustering of pages to find authoritative pages to improve retrieval performance.

C. *Web Usage Mining*

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site [5]. The strategic goals of Web usage mining are prediction of the user's behaviour within the site, comparison between expected and actual Web site usage, and adjustment of the Web site with respect to the users' interests.

The fist type of source data for web usage mining is Web Server Data. When web users interact with a site, their behaviour is recorded in web server logs on web server. These log files may contain valuable information characterizing the users' experience in the site. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users. In addition, since in a medium size site log files amount to several megabytes a day, there is a necessity of techniques and tools to help take advantage of their content. Application Level Data is another source for web usage mining. With this type of data it is possible to record various kinds of events in an application. These data is used for generating histories about selected special events. The data in this category can be divided into three categories based on the source of its collection: on the server side, the client side, and the proxy side.

III. Pattern Discovery Techniques

Pattern discovery techniques involve algorithms to discover interesting patterns from web data. Once user transactions or sessions have been identified, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst. Some of these discovery techniques are discussed below.

A. *Path Analysis*

Graph models are most commonly used for Path Analysis. In the graph models, a graph represents some relation defined on Web pages and each tree of the graph represents a web site. Each node in the tree represents a web page and edges between trees represent the links between web sites and the edges between nodes inside a same tree represent links between documents at a web site. When path analysis is used on the site as a whole, this information can offer valuable insights about navigational problems. Most graphs are involved in determining frequent traversal patterns and more frequently visited paths in a web site. For Example: What paths do users traversal before they go to a particular URL? Examples of information that can be discovered through path analysis are:

- 69% of clients who accessed /company/products/ order.asp by starting at /company and proceeding through /company/whatsnew.html, and /company/ products/sample.html ;

- 58% of clients left the site after four or less page references.

The first rule tells us that 69% of visitors decided to make a purchase after seeing the sample of the products. The second rule indicates an attrition rate for the site. Since many users don't browse further than four pages into the site, it is tactful to ensure that most important information for example product sample, is contained within four pages of the common site entry points.

B. *Association Rules*

Predict the association and correlation among set of items where the presence of one set of items in a transaction implies with a certain degree of confidence the presence of other items. That is, it can discover the correlations between pages that are most often referenced together in a single server session/user session. It can provide the information: What are the set of pages frequently accessed together by web users? What page will be fetched next? What are paths frequently accessed by web users?. Implement association rules to on-line shopper can generally find out his/her spending habits on some related products [3]. For example, if a transaction of an on-line shopper consists of a set of items, while each item has a separate URL. Then the shopper's buying pattern will be recorded in the log file, and the knowledge mined from which, can be the form like the following:

- 35% of clients who accessed the web page with URL /company/products/bread.html, also accessed /company/ products/milk.htm.

- 50% of clients who accessed /company/announcements/ special.html, placed an online order in /company/ products/products1.html

C. *Sequential Patterns*

Sequential patterns discovery is to find the inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set [9]. Web log files can record a set of transactions in time sequence. Using sequential pattern discovery, useful user trends can be discovered, predictions concerning visit pattern can be made, website navigation can be improved and adopt web site contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles. The sequential patterns can be discovered as the following form:

- 50% of client who bought items in /pcworld/computers/, also placed an order online in /pcworld/accessories/ within 15 days

D. *Decision Trees*

A decision tree is essentially a flow chart of questions or data points that ultimately leads to a decision [10]. For example, a car-buying decision tree might start by asking whether you want a 2008 or 2009 model year car, then ask what type of car, then ask whether you prefer power or economy, and so on. Ultimately it can determine what might be the best car for you. Decision trees systems are incorporated in product-selection systems offered by many vendors. They are great for situations in which a visitor comes to a Web site with a particular need. But once the decision has been made, the answers to the questions contribute little to targeting or personalization of that visitor in the future.

E. Clustering

Clustering identifies visitors who share common characteristics. After you get the customers'/visitors' profiles, you can specify how many clusters to identify within a group of profiles, and then try to find the set of clusters that best represents the most profiles [9]. Besides information from Web log files, customer profiles often need to be obtained from an on-line survey form when the transaction occurs. For example, you may be asked to answer the questions like age, gender, email account, mailing address, hobbies, etc. Those data will be stored in the company's customer profile database, and will be used for future data mining purpose. An example of clustering could be:

- 50% of clients who applied discover platinum card in /discovercard/customerService/newcard, were in the 25-30 age group, with annual income between $40,000 – 50,000.

Clustering of client information can be used on the development and execution of future marketing strategies, online and/or off-line, such as automated mailing campaign.

F. Grouping

Users usually can draw higher-level conclusions by grouping similar information. For example, grouping all Netscape browsers together and all Microsoft browsers together will show which browser is more popular on the site, regardless of minor versions. Similarly, grouping all referring URLs containing the word "Yahoo" shows how many visitors came from a Yahoo server. For example: http://search.yahoo.com/bin/search?p=Web+Miners

G. Filtering

Simple reporting needs require only simple analysis systems. However, as the company's Web becomes more integrated with the other functionality of the company, for example, customer service, human resources, marketing activity, analysis need to rapidly expand. For example, the company launches a marketing campaign. Print and television ads now are designed to drive consumers to a Web site, rather than to call an 800 number or to visit a store. Consequently, tracking online marketing campaign results is no longer a minor issue but a major marketing concern.

Often it's difficult to predict which variables are critical until considerable information has been captured and analysed. Consequently, a Web traffic analysis system should allow precise filtering and grouping information even after the data has been collected. Systems that force a company to predict which variables are important before capturing the data can lead to poor decisions because the data will be skewed toward the expected outcome. Filtering information allows a manager to answer specific questions about the site. For example, filters can be used to calculate how many visitors a site received this week from Microsoft. In this example, a filter is set for "this week", and for visitors that have the word "Microsoft" in their domain name e.g.proxy12.microsoft.com. This could be compared to overall traffic to determine what percentage of visitor's work for Microsoft.

H. Dynamic Site Analysis

Traditional Web sites were usually static HTML pages, often hand-crafted by Webmasters. Today, a number of companies, including Microsoft, make systems that allow an HTML file to be dynamically created around a database. This offers advantages like, included centralized storage, flexibility, and version control. But it also presents problems for some Web traffic analysis because the simple URLs normally seen on Web sites may be replaced by very long lines of parameters and cryptic ID numbers. In such systems, query strings typically are used to add critical data to the end of a URL usually delimited with a "?". For example, the following referring URL is from Netscape Search:

- http://search.netscape.com/cgi-in/search?search= Federal+Tax+Return+Form&cp=ntserch

By looking at the data after the "?" we see that this visitor searched for "Federal Tax Return Form" on Netscape before coming to our site. Netscape encodes this information with a query parameter called "search" and separates each search keyword with the "+" character. In this example, "Federal," "Tax," "Return" and "Form" each is referred to as parameter values. By looking at this information, companies can tell what the visitor is looking for. This information can be used for altering a Web site to ensure that information visitors are looking for is readily available, and for purchasing keywords from search engines.

I. Cookies

Cookies usually are randomly assigned IDs that a Web server gives to a Web browser the first time that the browser connects to a Web site. On subsequent visits, the Web browser sends the same ID back to the Web server, effectively telling the Web site that a specific user has returned. Cookies are independent of IP addresses, and work well on sites with a substantial number of visitors from ISPs. Authenticated usernames even more accurately identify individuals, but they require each user to enter a unique username and password, something that most Web sites are unwilling to mandate. Cookies benefit Web site developers by more easily identifying individual visitors, which results in a greater understanding of how the site is used. Cookies also benefit visitors by allowing Web sites to recognize repeat visits. For example, Amazon.com uses cookies to enable their “one-click” ordering system. Since Amazon already has your mailing address and credit card on file, you don’t re-enter this information, making the transaction faster and easier. The cookie does not contain this mailing or credit card information; that information typically was collected when the visitor entered it into a form on the Web site. The cookie merely confirms that the same computer is back during the next site visit.

If a Web site uses cookies, information will appear in the cookie field of the log file, and can be used by Web traffic analysis software to do a better job of tracking repeat visitors. Unfortunately, cookies remain a misunderstood and controversial topic. A cookie is not an executable program, so it can’t format your hard drive or steal private information. Modern browsers have the ability to turn cookie processing on or off, so users who chose not to accept them are accommodated.

IV. Issues of Web Mining in E-Commerce

There are some issues that need to be discussed in order to apply the web mining in e-commerce.

- Certain important implementation parameters in retail e-commerce sites like the automatic time outs of user sessions due to perceived inactivity at the user end need to be based not purely on data mining algorithms, but on the relative importance of the users to the organization [1]. It should not turn out that large clients are made to lose their shopping carts due to the time outs that were fixed based on a data mining of the application logs.
- Generating logs for several million transactions is a costly exercise. It may be wise to generate appropriate logs by conducting random sampling, as is done in statistical quality control. But such a sampling may not capture rare events, and in some cases like in advertisement referral based compensations, the data capture may be mandatory. Techniques thus need to be in place that can do this sampling in an intelligent fashion.
- Designing user interface forms needs to consider the data mining issues in mind. For instance, disabling default values on various important attributes like Gender, Marital status, Employment status, etc., will result in richer data collected for demographical analysis. The users should be made to enter these values, since it has found that several users left the default values untouched [2].
- Mining data at the right level of granularity is essential. Otherwise, the results from the data mining exercise may not be correct.
- Collecting data at the right level of abstraction is very important. Web server logs were originally meant for debugging the server software. Hence they convey very little useful information on customer related transactions. Approaches including sessionising the web logs may yield better results. A preferred alternative would be having the application server itself log the user related activities. This is certainly going to be richer in semantics compared to the state-less web logs, and is easier to maintain compared to state-full web logs [11][1].

V. Application of Web Mining in E-commerce

A. *Customer Attraction*

The two essential parts of attraction are the selection of new prospective customers and the acquisition of selected potential candidates. One marketing strategy to perform this exercise is to find common characteristics in already existing visitors’ information and behaviour for the classes of profitable and non-profitable customers [7]. These groups are then used as labels for a classifier to discover Internet marketing rules, which are applied online on site visitors. Depending on the outcome, a dynamically created page is displayed, whose contents depends on found associations between browser information and offered products / services. The three classification labels used were ‘no customer’, that is browsers who have logged in, but did not purchase, ‘visitor once’ and ‘visitor regular’. An example rule is as follows.

if Region = IRL and
Domain1 IN [uk, ie] and
Session > 320 Seconds
then VisitorRegular
Support = 6.4%; Confidence = 37.2%

This type of rule can then be used for further marketing actions such as displaying special offers to first time browsers from the two mentioned domains after they have spent a certain period of time on the shopping site.

B. *Customer Retention*

Customer retention is the step of managing the process of keeping the online shopper as loyal as possible. Due to the non-existence of physical distances between providers, this is an extremely challenging task in electronic commerce scenarios. One strategy is similar to acquisition that is dynamically creating web offers based on associations. However, it has been proven more successful to consider associations across time, also known as sequential patterns. The discovered sequence can be used to display special offers dynamically to keep a customer interested in the site, after a certain page sequence with a threshold support and / or confidence value has been visited.

C. *Cross-Sales*

The objective of cross-sales is to diversify selling activities horizontally or vertically to an existing customer base. Traditional generic cross-sales methodology has been adopted in order to perform the given task in an electronic commerce environment. For discovering potential customers, characteristic rules of existing cross-sellers had to be discovered, which was performed through the application of attribute-orientated induction [10]. For a scenario in which the product CD is being cross-sold to book sellers, an example rule is:

if Product = book then
Domain1 = uk and
Domain2 = ac and
Category = Tools
Support = 16.4%; Interest = 0.34

Deviation detection is used to calculate the interest measure and to filter out the less interesting rules. The entire set of discovered interesting rules can then be used as the model to be applied at run-time on incoming actions and requests from existing customers.

D. Improve the e-commerce web site design

Attractiveness of the site depends on its reasonable design of content and organizational structure. Web Mining can provide details of user behaviour, providing web site designers basis of decision making to improve the design of the site [6].

VI. Conclusion

The growth of World Wide Web and technologies has made business functions to be executed fast and easier. As large amount of transactions are performed through e-commerce sites and the huge amount of data is stored, valuable knowledge can be obtained by applying the Web Mining techniques. Using Web Mining, companies can understand customer behaviour, improve customer services and relationship and measure the success of marketing efforts. In this paper, we have discussed web mining in e-commerce, categories of web mining, pattern discovery, issues and application of web mining in e-commerce. The extension in web mining research will lead to success of e-commerce sites and also it will improve the services for customers.

References

[1] Hamid Rastegari, Mohd Noor Md. Sap, “Data Mining and E-Commerce: Methods, Applications and Challenges”, 2008.

[2] R. Kohavi, "Lessons and Challenges from Mining Retail E-Commerce Data," 2004.

[3] Penelope Markellou, Ioanna Mousourouli, Spiros Sirmakessis, Athanasios Tsakalidis, “Personalized E-commerce Recommendations”, Proceedings of IEEE International Conference on e-Business Engineering, 2005.

[4] G. Chang, M. J. Healy, J. A. M. McHugh, and J. T. L. Wang, “Mining the World Wide Web: An Information Search Approach”, Kluwer Academic Publishers, 2001.

[5] Y. Fu and M. Shih, “A Framework for Personal Web Usage Mining”, International Conference on Internet Computing, 2002, pages 595-600.

[6] TIAN Meirong, CHEN Xuedong, “Application of Agent-based Web Mining in E-business”, Second International Conference on Intelligent Human-Machine Systems and Cybernetics, 2010.

[7] Pradnya Purandare, “Web Mining: A Key to Improve Business On Web”, IADIS European Conference Data Mining, 2008.

[8] LIU Kainan, ZHANG Fengyan, PAN wumin, “Research on the Data Mining Technology of the E-Commerce Based on the Interest”, Second International Conference on Computational Intelligence and Natural Computing, 2010.

[9] Manoj Pandia, Subhendu Kumar Pani, Sanjay Kumar Padhi, Lingaraj Panigrahy, R.Ramakrishna, “A Review Of Trends In Research On Web Mining”, International Journal of Instrumentation, Control & Automation, Volume 1, Issue 1, 2011, pages 37-41.

[10] Joan Anderson, “Enhanced Decision Making using Data Mining: Applications for Retailers”, Journal of Textile and Apparel, Technology and Management, Volume 2, Issue 3, 2002.

[11] Kohavi R., “Mining e-commerce data: The good, the bad, and the ugly”, In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pages 8-13.