International Journal of Basic and Applied Research
Vol. 02 Issue-2 (December 2012) (88-96)
ISSN - 2249-3352

Web Usage Mining-Pattern Discovery & Its Applications

Deepti V.Patange

Department of Computer Science Arts, Science & Commerce College
Chikhaldara District-Amravati, Maharashtra

Email : deepti_pethkar@rediffmail.com

Key-words:
Web Usage Mining,
Personalization,
Pattern Discovery

Abstract: With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses. A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future.

Web usage mining is an application of data mining technology to mine the data of the web server log file. It can discover the browsing patterns of user and some kind of correlations between the web pages. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. Web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern.

1. INTRODUCTION

As in classical data mining, the aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. There has been huge interest towards web mining. In web mining, dataset is the huge web data. Web data contains different kinds of information, including, web documents data, web structure data, web log data, and user profiles data. Two different approaches are proposed on the definition of web mining. One approach is process-based and the other is data-based. Data-based definition is more widely accepted today. In this perspective, web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process.

Web Mining [1,2] can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories. A figure depicting the taxonomy is shown in Figure 1

Fig. 1 Taxonomy of Web mining

2. Why Web Usage Mining?

In this paper, we will emphasize on Web usage mining. Reasons are very simple: 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. The growth of some E-businesses is astonishing, considering how E-commerce has made Amazon.com become the so-called "on-line Wal-Mart". Unfortunately, to most companies, web is nothing more than a place where transactions take place. They did not realize that [3,4]as millions of visitors interact daily with Web sites around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on.

3. How to perform Web Usage Mining?

Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data (as discussed below). Web server log files were used initially by the webmasters and system administrators for the purposes of "how much traffic they are getting, how many requests fail, and what kind of errors are being generated", etc. However, Web server log files can also record and trace the visitors' on-line behaviors. For example, after some basic traffic analysis, the log files can help us answer questions such as "from what search engine are visitors coming? What pages are the most and least popular? Which browsers and operating systems are most commonly used by visitors?" Web log file is one way to collect Web traffic data. The other way is to "sniff" TCP/IP packets as they cross the network, and to "plug in" to each Web server.

After the Web traffic data is obtained, it may be combined with other relational databases, over which the data mining techniques are implemented. Through some data mining techniques such as association rules, path analysis, sequential analysis, clustering and classification, visitors' behavior patterns are found and interpreted.

The above is the brief explanation of how Web usage is done. Most sophisticated systems and techniques for discovery and analysis of patterns can be placed into two main categories, Pattern Analysis Tools and Pattern Discovery Tools, as discussed below in detail.

3.1 Pattern Analysis Tools

Web site administrators are extremely interested in questions like "How are people using the site?" "Which pages are being accessed most frequently?", etc. These questions require the analysis of the structure of hyperlinks as well as the contents of the pages. The end products of such analysis might include:

1. the frequency of visits per document,
2. most recent visit per document,
3. who is visiting which documents,
4. frequency of use of each hyperlink, and
5. most recent use of each hyperlink.

The techniques of Web usage patterns discovery, such as association, path analysis, sequential patterns, etc. (will be illustrated below in detail.

The common techniques used for pattern analysis are visualization techniques,[6] OLAP techniques, Data & Knowledge Querying, and Usability Analysis. However, this paper mainly focuses on the Pattern Discoveries, and the Pattern Analysis will not be discussed further in detail.

3.2 Pattern Discovery Tools

Pattern Discovery Tools implement techniques from data mining, psychology, and information theory on the Web traffic data collected.

3.2.1 Data Pre-processing

Portions of Web usage data exist in sources as diverse as[4,5,7] Web server logs, referral logs, registration-files and index server logs. This information needs to be integrated to form a complete data set for data mining.

However, before the integration of the data, Web log files need to be [8] cleaned/filtered, using techniques like filtering the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units.

Filtering the raw data to eliminate irrelevant items is important for web traffic analysis. Elimination of irrelevant items can be accomplished by checking the suffix of the URL name, [9] which tells you what format these kind of files are. For example, the embedded graphics can be filtered out from the Web log file, whose suffix is usually the form of "gif", "jpeg", "jpg", "GIF", "JPEG", "JPG", can be removed.

The next step is to integrate data from all sources to form a visitor profile data. Or we can say, the data in registration files (mainly visitors' demographic and household information) can be appended to log and forms data. The figure gives an example of data integration.

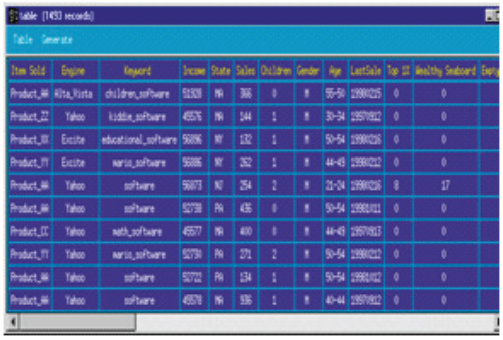

Figure 2.Integrate Data to form a visitors profile data

3.3 Pattern Discovery Techniques

3.3.1 *Converting IP addresses to Domain Names*

Every visitor to a Web site connects to the Internet through an IP address (for example, 198.227.55.153). Every IP address has a corresponding domain name, and these are linked through the Domain Name System (DNS). DNS can convert a domain name that a visitor entered in [10] Web browser into a corresponding IP address. A visitor's IP address can be converted into a domain name by using the DNS system in reverse, called a reverse DNS lookup.

You can hardly mine any knowledge merely from an IP number. However, if you convert the IP number into the domain name, some knowledge can be discovered. For example, you can estimate where visitors live by looking at the extension of each visitor's domain name, such as .ca (Canada); .au (Australia); cn(China), etc.

3.3.2 *Converting File Names to Page Titles*

A well-designed site will have a title (between <title> and </title>) for every page. Rather than simply report the file names (URL) requested, a good system should look at these files and determine their titles. Page titles are much easier to read than URLs, so a good system should show page titles on reports in addition to URLs.

3.3.3 *Path Analysis*

Graph models are most commonly used for Path Analysis. In the graph models, a graph represents some relation defined on Web pages (or web), and each tree of the graph represents a web site. Each node in the tree represents a web page (html document), and edges between trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site. When path analysis is used [11,12] on the site as a whole, this information can offer valuable insights about navigational problems. Examples [13] of information that can be discovered through path analysis are:

- 78% of clients who accessed */company/products/order.asp* by starting at */company* and proceeding through */company/whatsnew.html*, and */company/products/sample.html*;
- 60% of clients left the site after four or less page references.

The first rule tells us that 78% 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 (product sample, for example) is contained within four pages of the common site entry points.

3.3.4 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

3.3.5 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,[14,15] 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 analyzed. Consequently,[16] 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.

Dynamic Site Analysis / Vignette StoryServer

Traditional Web sites were usually static HTML pages, often hand-crafted by Webmasters. Today, a number of companies, including Vignette and 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 [16] 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.

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 a 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.

3.3.6 Association Rules

Implement association rules to on-line shopper can generally find out his/her spending habits on some related products.[17,18] 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:

- 30% of clients, who accessed the web page with URL /company/ products/ bread.html, also accessed /company /products/ milk.htm.
- 40% of clients who accessed /company/announcements/special.html, placed an online order in /company/products/products1.html

Another example of association rule shown below in figure 3 is the linked associations between online products and search keywords. It measures the association between the keywords used to search and the different products actually sold. This form of report can also be achieved by *Dynamic Site Analysis / Vignette StoryServer*mentioned above.

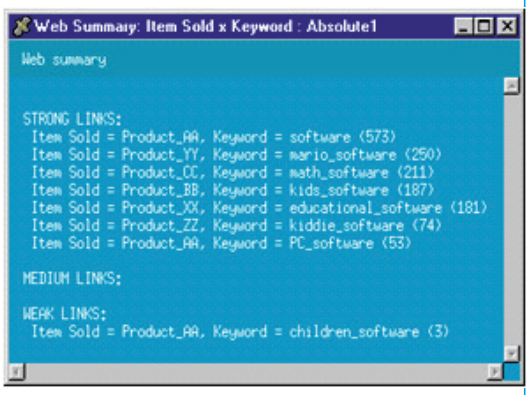

Figure 3.Linked Association between online products & search keywords

3.3.7 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. Web log files can record a set of transactions in time sequence. If the web-based companies can discover the sequential patterns of the visitors, the companies can predict users' visit patterns and target market on a group of users. 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

3.3.8 *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.

Besides information from Web[17],[19] 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.

3.3.9 *Decision Trees*

A decision tree is essentially a flow chart of questions or data points that ultimately leads to a decision. For example, a car-buying decision tree might start by asking whether you want a 1999 or 2000 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[20] 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.

4. Web Mining Applications

Web mining extends analysis much further by combining other corporate information with Web traffic data. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing, which answers complex questions such as:

- Of the people who hit our Web site, how many purchased something?
- Which advertising campaigns resulted in the most purchases, not just hits?
- Do my Web visitors fit a certain profile? Can I use this for segmenting my market?

Practical applications of Web mining technology are abundant, and are by no means the limit to this technology. Web mining tools can be extended and programmed to answer almost any question.

Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly. Also, the company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.

For example, the company may have a list of goals as following:

- Increase average page views per session;
- Increase average profit per checkout;
- Decrease products returned;
- Increase number of referred customers;
- Increase brand awareness;
- Increase retention rate (such as number of visitors that have returned within 30 days);
- Reduce clicks-to-close(average page views to accomplish a purchase or obtain desired information);
- Increase conversion rate (checkouts per visit).

The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement. This procedure is an on-going continuous process.

Next, we will give some examples on Web Mining applications.

4.1 Measuring Return of Online Advertising Campaigns

As online advertising banners become more popular, companies using them accurately measure overall return on advertising investment. This benefits both advertisers and sites running ads because it allows advertising rates to vary according to their success. Proper measurement of advertising reports centers on two specific areas:

Quantity: How many impressions were delivered for each ad banner and page, and how many people clicked on each ad? These are usually reported as *impressions* and *click-throughs*.

Quality: Of people who clicked on an ad banner, how many actually purchased? This return is best measured by subtracting advertising expenses from the resulting revenue.

For companies offering ad space on their site, reporting ad impressions and click-through rates for any page running advertisements is important. For companies running banner ads on other sites, prospect quality can be measured. A manager should evaluate both the effectiveness of individual ad banners and the effectiveness of each Web page with an [23]. By combining these, an advertiser optimizes his or her advertising by selecting the best combination of ad banner and Web page for additional ad placements. The following report

Table 1.Highest Click-Through Rates for Each Page

Page Name	Ad Name	Impressions	Click-Throughs	Click Through Rate	Cost
Front Page/default.htm	Mustang	34,100	21,00	6.2%	$3,410
	Sebring	34,600	1,400	4.0%	$3,460
	Corvette	92,100	3,100	3.4%	$9,210
	Intrigue	64,100	2,100	3.3%	$6,410
	Camaro	93,700	1,500	1.6%	$9,370
Classifieds/class.html	Corvette	9,800	300	3.1%	$980
	Intrigue	10,000	200	2.0%	$1,000
Hot Topics/hotnews.asp	Mustang	3,400	1,200	35.3%	$340
	Corvette	3,300	200	6.1%	$330
	Sebring	5,900	200	3.4%	$590
	Camaro	6,200	200	3.2%	$620

gives an example of a car site and the most effective ads for each page.

4.2 Measuring Return of E-Mail Campaigns

Combining Web traffic tools with an e-mail merging program is one of the best ways to maximize return on marketing e-mails. Custom URLs or query strings are assigned to each prospect. When the prospect reads the message and clicks the URL, the Web traffic analysis program determines who the visitor is and begins the appropriate sales process. For example: http://www.company.com /default.htm?Visitor=e-mail_address

When this URL is clicked, the unique identifier *(e-mail_address)* is passed to the Web site. By using a filter based on the query string, it is possible to measure the best leads. By linking the e-mail_address [24] to a customer information database, sales personnel can receive reports showing contact names, phone numbers, and more. Results measured in dollars can also be calculated by linking to a marketing database.

4.3 Market Segmentation

When combined with a profiling system, Web mining can perform market segmentation. This allows Web marketers to target campaigns and messages to specific groups. For example,[23] an online music company using a profiling system could create reports showing the differences in browsing behavior based on age ranges. They might find that most of their actual purchasers are in their 20's. An understanding of what information was attractive to other visitors would be valuable in designing a Web site to appeal to a wider audience. This information could be used to expand content and direct visitors to the right place. For example:

The above just a few sample applications of Web mining. As we said before, the practical applications of Web mining are abundant. Web mining is not exclusively implemented in the Internet, it can also implemented in Intranet (among the users within the company, mainly employees) and Extranet (suppliers and customers with EDI connection). [24] With the Web mining on Intranet and Extranet, company can achieve resource optimization within the organization, and improve customer service and/or supply train management with the suppliers (upstream) as well as with the customers (downstream).

Table 2.Name Age Range Requests

Age Ranges per Page		
Shows the age ranges, in 10 year increments, of visitors to each page.		
Page Name	**Age Range**	**Requests**
Home Page	0: 9	1
	10: 19	3
	20: 29	23
	30: 39	13
	40: 49	11
	50: 59	8
Product Page	0: 9	1
	10: 19	4
	20: 29	263
	30: 39	141
	40: 49	23
	50: 59	71
Customer Support Page	20: 29	21
	30: 39	14
	40: 49	8
	50: 59	8

5. Summary

It is a revolution that the Internet has grown from a simple search tool to a gold mine. Companies find a new and better way to do business: E-commerce through the Internet. However, E-business cannot just build a web site and then sit back and reap the benefits, which, in most cases, is fruitless. 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.

REFERENCES

[1] R. Kosala, H. Blockeel,SIGKDD Explorations 2(1), ACM, July 2000.
[2] Ramakrishna, M.T. Gowdar, L.K. Havanur, M.S. Swamy, International Conference on Data Storage and Data Engineering (DSDE), pp.187 – 191, 2010.
[3] G. Castellano, A. M. Fanelli, M. A. TorselloProceedings of the 6th WSEAS Int. Conf. on Simulation, Modelling and Optimization 2006
[4] J. Srivastava, R. Cooley, M. Deshpande, P.N. Tan. SIGKDD Explorations, Vol1, Issue 2, 2000.
[5] Jianxi.Zhanga, PeiyingZhaob, LinShanga, LunshengWangaISECS International Colloquium on Computing, Communication, Control, and Management,2009
[6] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, User Modeling and User- Adapted Interaction, Vol. 13, No. 4, pp. 311-372, 2003
[7] M .Kitsuregawa, M. Toyoda, I. Pramudiono, In Proceedings of the 13th Australasian Database Conference ADC (02), Melbourne, Australia, 5, pp. 3–10,2002,.
[8] Shinde S.K Advanced computer theory and engineering, ICACTE pp 973-977, 2008
[9] K. Etminani, A. Delui, IEEE Transcations First International Conference on Networked digital technologies 2009.
[10] A Joshi, R Krishnapuram, SIGMOD 1998.
[11] Robert Cooley, JaideepSrivastava , URLs 1999
[12] Joachims, T., Freitag, D., Mitchell, T., In Proceedings of the Int. Joint Conference in AI (IJCAI97), August 1997.
[13] Alex G. Büchner, Maurice D. Mulvenna Discovering Internet marketing intelligence through online analytical web usage mining, Volume 27 Issue 4, Pages 54 – 61 Dec 1998.
[14] RenataIvancsy And Ferenc Kovacs (2006)" Madrid, Spain, February 15-17, pp237-242 2006
[15] R. T. Ng and J. Han, "Clarans: IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 5, pp. 1003–1016, 2002.
[16] S. Guha, R. Rastogi, and K. Shim, Washington, USA (L. M. Haas and A. Tiwary, eds.), pp. 73–84, ACM Press, 1998.
[17] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," pp. 103–114, 1996.
[18] G. Sheikholeslami, S. Chatterjee, and A. Zhang,Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pp. 428–439, 24–27 1998.
[19] M.Ester, H.-P. Kriegel, J. Sander, and X. Xu, KDD, pp. 226–231, 1996.
[20] Maofu Liu Yanxiang He HuijunHuWebProceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization 2006
[21] Steve Russell, Intelligence for Business at E-Speed
[22] Database Access Over the Web: Extending the Wire, by Dr. Larry R. Harris
[23] Data Mining and the Web: What They Can Do Together, by Mary Garvin
[24] Web Mining White paper - Driving Business Decisions in Web Time