Saša Bošnjak
Mirjana Marić
Zita Bošnjak

The Role of Web Usage Mining in Web Applications Evaluation

Article Info:

Management Information Systems,
Vol. 5 (2010), No. 1,
pp. 031-036

Received 27 Jun 2009
Accepted 21 October 2009

UDC 005.21:004.738.5

Summary

The role of Web applications in corporate business has changed due to strong market competition and improved clients' negotiation power, imposing a new approach to their quality evaluation, in a sense that some analysis should be made prior to the implementation phase in order to reduce errors and inconsistencies in application design, while after the implementation, visiting scenarios and visitors' usage habits should be analyzed. The most common approach to the later task is to mine web access logs enriched by semantic information - web usage mining, in order to discover patterns hidden in data obtained through interaction of users on the web. In the paper we firstly provide a brief overview of data preprocessing, pattern discovery and pattern analysis steps of web mining and of most common pattern discovery methods. The remaining section demonstrates a practical example of Web site evaluation.

Key words

Web applications, web usage analysis, web usage mining, WebMl, web ratio.

Introduction

In the early stages of Web development, it was common to build Web applications in an *ad hoc* manner, without following any strict development methodology. Such an approach did not offer a strong foundation for application quality measurement and control. The intensified e-business applications development, the stronger market competition and especially the stronger negotiating power of clients have made the role of Web applications for overall business more significant, imposing a need for more reliable development process and therefore more accurate evaluation possibilities. Tools for complex Web applications generation, such as Microsoft's Active Server Pages, support just one phase in software development process and hence are not sufficient for achieving high quality in Web applications development. Recently the conceptual modeling of Web applications has been introduced as a new paradigm.[1] Generated conceptual schemes, an output from the Web application design phase, are investigated twofold: (1) a quality analysis is performed prior to the implementation of Web application, in order to identify the errors and inconsistencies in the application design that could jeopardize its usefulness; (2) a quality analysis is performed after the implementation of Web application, by measuring its usefulness, by analyzing visiting scenarios and visitors' usage habits, by means of diversified data analysis and data mining techniques. Both approaches have the same goal: to exploit the conceptual application scheme obtained in the design phase, to interpret the qualitative evaluation results and transform them into corrective actions applicable to previous application design and transferable into application code. The *a posteriori* analysis includes the Web Usage Analysis (WUA) and the Web Usage Mining (WUM). WUA is based on conceptual logs and semantically enriched log data collected in runtime. The results of the analysis are reports on content accessibility and navigation paths visitors are using. WUM operates on conceptual logs by means of data mining techniques for discovering interesting visitors' behavior, not predicted in beforehand by the application designer. The revealed patterns in visiting habits usually result in application interface revision.

In this article we are interested in Web usage mining that uses web data sources in order to discover hidden knowledge about users and their behavior on the Web. Such knowledge, if taken advantage of, brings to organization nothing than benefits and leads directly to profit increase. Site modification, business intelligence, system improvement, personalization and usage characterization are the areas in which the potentials of Web usage mining have been recognized and extensively used (Cooley, Mobasher, & Srivastava, 1999).

[1] One tool that supports the overall Web application development cycle is Web Ratio, with *WebMl* design language and incorporated design methods of conceptual modeling.

1. Web usage mining frameworks

With about 30 million new web pages posted every day, the WWW is the largest, most used, and most important knowledge source and the most perspective marketplace. In order to successfully retain users in this rapidly developing environment, a web site must be built in such a way that supports user personalization. To achieve this, an organization can keep track of user activities while browsing their web sites. Although there are many tools that help analyze this data using some of the web statistics methods, they provide sufficient information only for the web site administrator (e.g. discovering part of a day with the most traffic, most frequently visited pages, etc.) and not for the designer. One of the ways to overcome this shortcoming is by applying data mining techniques on the Web. This process is referred to as web mining. There are three constituents of web mining: content mining, usage mining and structure mining. As most web servers keep logs, the most common data sources are Web access logs (clikcstream data).

Since web usage mining is a relatively new area of data mining, many authors (and software companies) have developed frameworks for it. Today there is a multitude of tools that support web usage mining based on various frameworks. The difference between frameworks ranges from slight to completely different philosophy.

One of the most used and referenced framework for web page personalization using web access was developed by Cooley et al. (1999). This framework follows the three-step process. In the first step, it was suggested to process the data not only from log data, but to use site topology and page classification (head, content, navigation, look up, personal) based on physical and usage characteristics, so afterwards this heuristics can be used to identify users and sessions. Then, data referring to sessions are transformed into transactions which represent clusters of page preferences for each user. Data cleaned and transformed in this way is then presented to some of the pattern discovery methods. The way this method can be used for web personalization is described in Cooley et al. (1999).

In another model, proposed by Zaini, web access logs are inserted into a multidimensional cube and then analyzed using OLAP and data mining techniques. In this approach we analyze data stored in fact and dimension tables (star schema). There are three kinds of fact tables: click table that stores data about each log entry, session facts (a user session is a set of pages visited by the user in one visit to a web page) and a subsession (sequence of clicks in a session).

Combining two approaches, Jespersen, Thorhauge, & Pedersen (2002) developed a hybrid approach to web usage mining, which combines the compact HPG (Hyper Probability Grammar) approach with the explicit OLAP approach. In this model, data is stored into a database through the XML and Quilt Query. The constraints for the analysis are built on the top of this database and data together with the constraints are used for modeling Hypertext Probabilistic Grammars, which are then mined using Breadth First Search (BFS) based algorithm for mining association rules.

2. The WUM process

We can identify three basic steps that the web usage mining process must follow. These steps are *data preprocessing, pattern discovery* and *pattern analysis.*

To successfully complete an analysis of a web site, we must obtain data suitable for data mining at the beginning of a process. Most of the authors in their papers agree that data preprocessing step is the most time-consuming step in web usage analysis projects (from 60 to 90 % of the time necessary for the completion of an entire project). The task of data preprocessing is to prepare the data for the application of some data mining algorithm.

After data has been preprocessed, it is ready for the application of knowledge extraction algorithms. When exposed to these algorithms, data in web access logs can be transformed into knowledge, most commonly, about association rules, sequential patterns and user clusters.

The last phase in WUM process is the analysis of the obtained results in order to distinguish trivial, useless knowledge from knowledge that could be used for Web site modifications, system improvement and/or Web personalization.

2.1. Data preparation

There are a number of sources for web usage mining, such as web access logs, cookies, data tags, login information, client or server side scripts, packet sniffing, etc. Each of these so-called e-sources has its advantages and disadvantages, but web access logs are the main sources for web usage mining. They are recorded in standard formats, usually the CLF – Common Log Format and ECLF – Extended Common Log Format by leading Web Server (Apache, IIS, Netscape). The structure of these log formats is defined by W3C. The CLF has the following 7 elements (shown in Table 1):

- *remotehost* - domain name or IP address
- *rfc931* - the remote logname of the user
- *Authuser* - user identification used in a successful SSL request
- *[date]* - the date and time of a request (e.g. day, month, year, hour, minute, second, zone)
- *"request"* - the request line exactly as it came from the client
- *status* - three-digit HTTP status code returned to the client (such as 404 for *Page not found*, or 200 for *Request fulfilled*)
- *bytes* - number of bytes returned to the client browser for the requested object ECFL has two additional elements:
- *referrer* - URL of the referring server and the requested file from a site
- *agent* - Browser and operating system name and version

Table 1 Examples of Common Log Format

Remotehost	rfc931	authuser	[date]	"request"	status	bytes
213.240.4.193	-	-	[04/Apr/2005:10:40:52 +0200]	GET /images/nastava%20color.jpg HTTP/1.1	200	13465
213.240.4.193	-	-	[04/Apr/2005:10:40:53+0200]	GET /images/zaglavlje.jpg HTTP/1.1	304	-
213.240.4.193	-	-	[04/Apr/2005:10:40:58+0200]	GET /raspored_ispita.htm HTTP/1.1	304	-
213.240.4.193	-	-	[04/Apr/2005:10:40:59+0200]	GET /images/ispit.jpg HTTP/1.1	304	-
213.240.4.193	-	-	[04/Apr/2005:10:41:02+0200]	GET /obavestenja.htm HTTP/1.1	304	-
213.240.4.193	-	-	[04/Apr/2005:10:41:02+0200]	GET /images/obavestenja.jpg HTTP/1.1	304	-
213.240.4.193	-	-	[04/Apr/2005:10:41:11+0200]	GET /obavestenja/naukaoradu.htm HTTP/1.1	200	18959
212.200.136.5	-	-	[04/Apr/2005:10:41:16+0200]	GET / HTTP/1.0	304	-
212.200.136.5	-	-	[04/Apr/2005:10:41:16+0200]	GET /images/zaglavlje.jpg HTTP/1.0	304	-
212.200.136.5	-	-	[04/Apr/2005:10:41:19+0200]	GET /images/efsuzgrada01.jpg HTTP/1.0	304	-
212.200.136.5	-	-	[04/Apr/2005:10:41:21+0200]	GET / HTTP/1.0	200	7295
212.200.136.5	-	-	[04/Apr/2005:10:41:23+0200]	GET /images/zaglavlje.jpg HTTP/1.0	200	18675
212.200.136.5	-	-	[04/Apr/2005:10:41:23+0200]	GET /images/efsuzgrada01.jpg HTTP/1.0	200	9843

This structure provides the basis for web usage mining. Although it is standardized, many problems are encountered when using clickstream data. The first task is to clean the data, i.e. to remove those log entries we do not need. Such entries are, for example, the "POST" method, error status entries (all log entries with status 4xx), bots and spider requests, requests for pictures, etc. These trivial data can make up to 50 % of the actual data size. After this is done, we need to identify the user (person who is visiting the web page), the pages he is visiting (set of web resources, which can be anything that has its identity, meaning that it can have an assigned URL (Cooley et al., 1999), and sessions (set of all pages visited by the user during one visit). The main problem associated with determining users, for sites that do not have logons or use https (second and third entries in web access logs), are proxies, DHCP servers and caching. If a user accesses the Internet through a company's or ISP proxy server, some of his requests are not logged by Web server since the proxy server contains some of frequent requests. DHCP servers make the task of user identification even more difficult, since the user's IP address is dynamically assigned each time he connects to the Internet. As a result, the same IP address in logs can represent various users and the same user can have more IP addresses. Caching web pages creates additional problems, since some of the pages are stored in memory and when, for example, the user clicks the back button on his browser, the request is not made to the server (the cached page is shown). To overcome the listed problems in data preprocessing, several techniques have been developed. The use of cookies to identify users is one of them. An advantage of cookies is their small size and their ability to store any desired string, while their disadvantage is that users can reject cookies through their web browser. Yet another method of data preprocessing is the so-called cache busting that prevents browsers or proxy servers from serving contents from their cache, so that a browser or a proxy server needs to fetch a fresh copy for each user request. To acquire more precise data about users, some web sites require logons or have a small JavaScript to identify users, or use data tags, while some servers use server-based scripts. All of this helps us to identify users, but there is still no definitive method for user or session identification. Cooley et al. (1999) follow two steps in the creation of a transaction model. In the first step, they group logs into clusters, so called LEGs (Log entry groups), which contain the user's IP, userid, and urls and times they have been requested. The distance which is used for creating these clusters is the time of the request (session timeout). The next step is to transform LEGs into an appropriate transaction model for the analysis. In their later work, the same authors expand this model by introducing a site topology to make transactions not only based on time, but on the maximum forward reference or reference length (Cooley et al., 1999). An approach of clustering logs into user sessions using the agglomerative clustering is described. The authors firstly define a distance based on overlapping of URLs, and then use a CARD (Competitive Agglomeration algorithm) to create user session clusters. Since their approach is based on heuristics, an application of soft computing techniques is well suited for user session identification. The i-Miner system uses the fuzzy C-means algorithm, optimized by the use of genetic algorithm, to make clusters from preprocessed data (i-miner).

Tanasa & Trousse (2004) give an overview of data preprocessing tasks through a description of several steps one must pass, and a comparison of various approaches for data preprocessing of web access logs for data mining.

2.2. WUM algorithms

Association rules are a data mining technique that searches for relationships between attributes in large data sets. They can be formally represented as:

$$X \rightarrow Y \qquad (1)$$

having X, Y ∈ D, D representing the set of all attributes, the so called *itemset* and $X \cap Y = \emptyset$.

If T denotes all transactions t, such that $t \in T$, and if there is an attribute X in transaction t, $X \subset t$, there is probably an attribute Y in t as well, $Y \subset t$ The possibility of this happening is called association rule confidence, denoted by *c* and measured as a percentage of transactions having Y along with X compared to the overall number of transactions containing X. Another important parameter describing the derived association rule is its support, denoted by *s*. It can be calculated as a percentage of transactions containing X and Y to overall number of transactions. These two metrics determine the significance of an association rule. Since the association rules tend to find relationships in large datasets, it would be very time and resource consuming to search for the rules among all data. Because of this each algorithm for discovering association rules begins with the identification of so called *frequent itemsets*. The most popular algorithms use two approaches for determining these itemsets. The first approach is BFS (breath-first search) and is based on knowing all support values of $(k-1)^{th}$ itemset before calculating the support of the k^{th} itemset. DFS (depth-first search) algorithms determine frequent itemsets based on a tree structure. The best known algorithms for mining association rules are Apriori, AprioriTID, STEM, DIC, Partition-Algorithm, Elcat, FP-grow, etc.

In web usage mining, association rules are used to discover pages that are visited together quite often. Knowledge of these associations can be used either in marketing and business or as guidelines to web designers for (re)structuring Web sites. Transactions for mining association rules differ from those in market basket analysis as they can not be represented as easily as in MBA (items bought together). Association rules are mined from user sessions containing remotehost, userid, and a set of urls. As a result of mining for association rules we can get, for example, the rule: $X,Y \rightarrow Z$ (c=85%, s=1%). This means that visitors who viewed pages X and Y also viewed page Z in 85 % (confidence) of cases, and that this combination makes up 1% of all transactions in preprocessed logs. In (Cooley et al., 1999) a distinction is made between association rules based on a type of pages appearing in association rules. They identify Auxiliary-Content Transactions and Content-only transactions. The second one is far more meaningful as association rules are found only among pages that contain data important to visitors.

Another interesting application of association rules is the discovery of so called *negative associations*. In mining negative association rules $(X \rightarrow \rceil Y)$ items that have less than minimum support support are not discarded. Algorithms for finding negative association rules can also find indirect associations.

Sequential patterns are another technique for pattern discovery commonly used for discovering knowledge in web access logs. Essentially sequential patterns differ from association rules because they consider the influence of time (timestamp). These timestamps are found in web access logs. Sequential patterns are trying to discover which items are followed by another set of items. For mining sequential patterns from web access logs it is required that each transaction contains the [date] field and a field that denotes the period of time for which we are mining sequential patterns. For example, 10% of visitors who visited page X followed up to page Y. This percentage is called support. Discovering sequential patterns can be used for predicting future visits and developing suitable Web site interface designs for them.

Clustering determines which elements in a dataset are similar. In web usages mining various clustering techniques are applied both for page clustering and user clustering. Page clustering tends to find information about similarities between web

Table 2 Association rules derived by market-basket analysis

		Support	Confidence
	Filter:	> 2.00	> 50.00
/obavestenja/inftehpitanja.htm	-> /obavestenja/infteh.htm	3.57%	96.59%
/Download/mikroekonomija/rezim_studiranja.htm	-> /Download/mikroekonomija/mikroekonomija.htm	4.07%	94.17%
/poslediplomske/konkurs2005.htm	-> /poslediplomske/pds_nastavni_plan.htm	2.10%	62.50%
/mapa_sajta.htm	-> /katedre.htm	2.10%	56.82%
/raspored_ispita/su_god_03.htm	-> /raspored_ispita.htm	3.49%	81.37%
/predavanja/ns1-zimski200506.htm	-> /nastava.htm	2.98%	76.34%

pages based upon visits. User clustering tries to discover groups of users having similar browsing patterns. Such knowledge is especially useful in E-commerce applications for inferring user demographics in order to perform market segmentation while in the evaluation of Web site quality this knowledge is valuable for providing personalized Web content to the users.

3. WUM case study

As a practical example of WUM, we have undertaken the analysis of the Web access logs of the Faculty of Economics Subotica Web site: www.eccf.su.ac.yu. The data originates from November 2005. We conducted the data mining process by a combination of data mining tools applied upon the relational database. In other words, we did not use any WUM tool, but used firstly the SQL Server 2000 relational database for storing the data collected from Apache Web server, and then we utilized the Association Rules module of PolyAnalyst software product. This tool offers great possibilities not just for Market Basket analysis, but for transactions analysis as well. The source data were in CLF format and amounted to 215.264 inputs. In the data preprocessing phase we had to erase all irrelevant inputs, as well as the demands for Web site versions in languages other than Serbian (English, German and Hungarian). Furthermore, we had eliminated all requests issued by search agents, the so called bots. The reason for this is that even though bots should search for rbot.txt files from server or have requests of type *head*, it is sometimes not the case. We had also discarded all input transactions exceeding the predefined number of requested pages of a Web site, as bots during Web search demand the majority of Web site pages. The remaining data set still consisted of 52.450 inputs, which is in accordance with the expectations of some 50% reduction in item number compared to the original data set. These preprocessed data were then grouped into sessions and further into transaction using the time interval of 15 minutes for each session. By following this grouping method we got 11.380 transactions, i.e. probable visits to our Website in November. Within the data set preprocessed in the above described way, we tried to discover some association rules that could help in the analysis of Web site navigation design. In this phase we used the PolyAnalyst software tool. Unfortunately, the limitation of the version of PolyAnalyst product available to us, of maximum of 10.000 inputs, has restricted our analysis: within this amount of input data, only 3.625 transactions could have been identified. Some association rules with the degree of confidence of 50% and the support degree of 2% are shown in Table 2.

The column with the heading Filter corresponds to the left-hand side of the association rule scheme. If we further analyze the results in the table, we can easily see that the first three rules refer to the start of colloquia on Information Technologies and Microeconomics subjects that are attended on the first year of studies. As on the first academic year there is the largest number of students, it is quite obvious that the degrees of confidence are among the highest. The third rule refers to the beginning of enrollment for postgraduate studies, which was in progress from October 15th till November 15th, 2005. The question that arises is: How to take advantage of these derived association rules? The through benefit in this case is the detected need for Faculty Web site modification in order to provide students with more efficient access to the information they are searching for. For example, in order to visit the Web page containing information on subject Microeconomics, a visitor (student) has to visit additional three pages he has probably no interest in. Having in mind the knowledge revealed by association rules, we cold modify the Web site in a way that the most frequently visited pages become more easily accessible.

4. Conclusion

This paper gives an insight into the possibility of merging data mining techniques with Web access logs analysis for achieving a synergetic effect of Web usage mining and its utilization in Web Applications Evaluation. The paper firstly describes the data preprocessing and pattern discovery steps, as two basic steps in the process of WUM, which Web designers should follow in knowledge extraction. Secondly it shows, on an illustrative example, how can hidden patterns, valuable for Web site designer, can be discovered from Web access logs data within the described framework and highlights the strong connection between Web site structure design and its quality. The selection of association rules discovery algorithm as a WUM technique for the described case study should by no means be understood as a suggestion that it is the best WUM algorithm, but as a convenient framework for our research. It is hard, if not impossible, to declare that one data mining algorithm is the best in general, because the possible outcomes of WUM process always depend on the problem in hand. Despite the difference in frameworks, knowledge hidden in clickstream data, discovered in WUM process, could and should be used for Web site design evaluation and further for undertaking corrective actions and making consequent improvements of the previous application design (redesign, personalization, etc.).

References

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference* (pp. 487-499). Santiago: Morgan Kaufmann.

Borges, J., & Levene, M. (1999). *Data mining of user navigation patterns.* Retrieved April 21, 2009, from Computer Science and Information Systems Birkbeck University of London: http://www.dcs.bbk.ac.uk/~mark/download/web_mining.pdf

Chelcea, S., DaSilva, A., Lechevallier, Y., Tanasa, D., & Trousse, B. (2005, November 30). *Benefits of InterSite Pre-Processing and Clustering Methods in E-Commerce Domain.* Retrieved April 21, 2009, from HAL :: Accueil: http://hal.archives-ouvertes.fr/docs/00/04/97/92/PDF/2005-DiscoveryChallenge_ChelceaEtAll.pdf

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and InformationSystems , 1* (1), 5-32.

Jespersen, S. E., Thorhauge, J., & Pedersen, T. B. (2002). *A Hybrid Approach to Web Usage Mining.* Retrieved April 22, 2009, from SpringerLink: http://www.springerlink.com/content/26rynqvgkhephq1x/fulltext.pdf

Menasalvas, E., Marban, O., Millan, S., & Pena, J. (2003). Intelligent Web mining. In P. S. Szczepaniak, J. Segovia, J. Kacprzyk, & L. A. Zadeh (Eds.), *Studies In Fuzziness And Soft Computing: Intelligent exploration of the web* (pp. 363-388). Heidelberg: Physica-Verlag GmbH.

Tanasa, D., & Trousse, B. (2004). Advanced Data Preprocessing for Intersites Web Usage Mining. *IEEE Intelligent Systems Magazine , 19* (2), 59-65.

Wang, X., Abraham, A., & Smith, K. (2005). Intelligent web traffic mining and analysis. *Journal of Network and Computer Applications* (28), 147-165.

Saša Bošnjak
University of Novi Sad
Faculty of Economics Subotica
Segedinski put 9-11
24 000 Subotica
Serbia
E-mail: bsale@ef.uns.ac.rs

Mirjana Marić
University of Novi Sad
Faculty of Economics Subotica
Segedinski put 9-11
24 000 Subotica
Serbia
Email: mprokic@ef.uns.ac.rs

Zita Bošnjak
University of Novi Sad
Faculty of Economics Subotica
Segedinski put 9-11
24 000 Subotica
Serbia
E-mail: bzita@ef.uns.ac.rs