Online Journal of Applied Knowledge Management
A Publication of the International Institute for Applied Knowledge Management

Volume 1, Issue 2, 2013

Association rules for improving website effectiveness: case analysis

Maja Dimitrijević, The Higher Technical School of Professional Studies, Novi Sad, Serbia, dimitrijevic@vtsns.edu.rs

Tanja Krunić, The Higher Technical School of Professional Studies, Novi Sad, Serbia, krunic@vtsns.edu.rs

Abstract

Association rule mining of the web usage log files can be used to extract patterns of a website visitors' behavior. This knowledge can then be utilized to enhance web marketing strategies or improve the web browsing experience. In this paper we apply association rule mining on the web usage log file of an educational institution. We use confidence and lift as the association rule interestingness measures and compare their values in two different time periods. We show how this comparison brings additional information about association rules discovered and helps a webmaster make more informed decisions about the website enhancements.

Keywords: web usage mining, association rules, interestingness measures, e-business

Introduction

With the intense growth of e-commerce in recent years, the use of intelligent data management strategies has become essential in the e-business community. The knowledge about client behavior may potentially be utilized to maximize on-line application effectiveness, increase client satisfaction, and competitiveness.

Association rule mining is a data mining method originally invented to extract patterns from transactional databases. Stated simply, an association rule is an implication in the form $X \rightarrow Y$, where X and Y are sets of items. Association rule mining identifies all such implications existing in a given data set, which satisfy certain constraints, such as minimal support and minimal confidence (Agrawal, Imielinski, & Swami, 1993).

An interestingness measure is used to assign a value to each association rule in order to distinguish those rules that are potentially most interesting to a data analyst in a given domain. It usually measures how tightly the itemsets X and Y are correlated. In addition to confidence, which is basically the conditional probability of Y given X, many other statistical measures have been used to determine rule interestingness (Geng & Hamilton, 2006; Tan & Srivastava 2004).

Association rule mining can be used to mine web usage log files and extract patterns about website visitor behavior. In this context a web resource is considered an item, while a website visitor session is considered a transaction of items (Anand et al., 2004; Kosala & Blockeel, 2000).

In this research we mine association rules from the web usage log files of a website of an educational institution. We propose to consider the changes of confidence levels of the association rules found in two different time periods. We discuss how useful the results are for a webmaster in order to make an action to increase the website effectiveness.

The rest of the paper is organized as follows. The *Related Work* section discusses related issues and approaches to web usage association rule mining. The section *Discovering Association Rules* explains our research method applied to a real life dataset. The section *Case discussion from the webmaster's perspective* shows a sample of web usage association rules found in our data and the related comments from a webmaster's perspective. Our conclusions are given in the *Conclusion* section.

Related Work

Typically association rule algorithms generate a huge number of rules, not all of which are truly interesting to a data analyst. Various methods have been proposed to prune the set of generated rules and discard irrelevant rules during the rule generation phase (Sahaaya & Malarvizhi, 2010). Balcazar (2010) presented a formal approach to determining a rule set basis and thus eliminate redundant rules. Although various pruning methods can significantly reduce the rule set size, it typically still remains huge.

An important issue in association rule mining is selecting the right interestingness measures (Huang & Xiangji, 2007). A recent study (Kannan & Bhaskaran, 2009) analyzes the distribution of rule clusters over various interestingness measures.

A discussion of the limitations of association rule mining is given in a study by Webb (2011). These include the over-generation of rules, many of which are false discoveries, discovery of rules that are hard to understand, and limitations imposed by the minimal support constraint. The authors propose an approach called "filtered top-k association discover" that alleviates some of these problems.

Association rule mining in the context of web usage data suffers from additional issues. Namely, web usage data is specific in the sense that it contains a large number of tightly correlated items (web resources or web pages) due to the link structure of a website. Web pages that are tightly linked together often co-occur in the same session, , is why the generated set of association rules contains a high number of so-called "hard" association rules that have very high confidence, but are not truly interesting to the user. To alleviate this problem, additional rule pruning methods need to be applied (Wang, 2005; Dimitrijevic & Bosnjak 2010).

A line of research deals with finding additional web usage association rules that may be missed during the classical association rule mining. For example, an approach that discovers association rules between web pages that rarely occur together, but with other pages, where they occur frequently, is taken by Kazienko (2009). This method is used to add new, so called "transitive" association rules to the typical association rule set, which brings new potentially useful information to a webmaster.

In this study, we propose to consider the changes of the association rule interestingness values in two different time periods. We discuss how this brings additional knowledge about the interestingness of web usage association rules to a webmaster. We apply a simple pruning

strategy in order to alleviate the problem of over-generation of not truly interesting rules. We conduct an empirical study on the website of an educational institution and present the results.

Discovering association rules

Data set

For the purpose of this research, we have conducted an empirical study on the website of an educational institution. The web usage log file mined contains information about all web requests made in November 2012. The raw web usage log file contains 649,749 web requests. We also used the results of the web usage log mining of the same educational institution website that contained all web requests made to the same website in the same month of 2010 (Dimitrijevic & Bosnjak, 2011). The raw web usage log file contained 397,741 web requests.

Research Method

Our web usage log file mining process consists of four phases.

Phase 1: Preprocessing

The purpose of the first phase is to clean the web usage log file and eliminate irrelevant entries. The log file is preprocessed using WumPrep tool (Dettmar, 2004) to exclude irrelevant entries and to group all entries into visitor sessions.

We refer to a session as a set of web resources requested during a website visit. When a website visitor browses through a website and then makes a pause and returns, her/his visit may be considered as one or two sessions. In this work we use the definition of a session as the set of web resources requested from the same IP address where the time between two consecutive requests does not exceed 5 minutes (Anand et al., 2004; Dettmar, 2004).

The resulting preprocessed files contain 16,637 visitor sessions made in November 2010 and 36,354 visitor sessions made in November 2012.

Phase 2: Association rule mining

We use our software (Dimitrijevic & Bosnjak, 2011) to mine association rules from the web usage log file. For the purpose of this study, we focus on the so called "short" rules that contain one url on each side of the rule. Those rules are easier to understand by the webmaster and more likely to cause a webmaster to act and change a website structure.

In order to select the most interesting association rules, we combine the confidence and lift interestingness measures. We sort the association rule set according to lift, while using a minimum confidence threshold to prune the non-interesting rules. We set the minimum support threshold value to 0.02 and the minimum confidence threshold value to 0.2.

Phase 3: Pruning association rule set

In order to minimize the number of "hard" association rules in our rule set, which result from the inter-connectedness of web pages through the website link structure and are not truly interesting to a webmaster, we apply a simple pruning methodology adopted from (Dimitrijevic & Bosnjak, 2011). All rules that contain directly linked pages are pruned out of the rule set. The term

"directly linked" is limited to the links in the body text of the page only and excludes the links through main menus.

Phase 4: Joining the rule sets

We join the rule sets in order to compare the rules found in November 2010 and in November 2012 and highlight confidence changes in the rules that appear in both sets. The file containing summarized information about rules from both rule sets and their confidence level changes is presented to the webmaster for further analysis and comments.

Dynamic Confidence

One of the main goals of this study was to investigate how the confidence of the discovered rules changes over time, and how that affects a webmaster when making decisions based on the knowledge contained in those rules. Therefore, we compared the rule sets generated in November 2010 and in November 2012 from the web usage log file of the educational institution. The website of the institution has undergone minimal changes within this period. Some pages have been added, but most of the old pages still exist, and their content had not changed. Hence most rules found in the old rule set were still there in the new set giving us the opportunity to compare the rules and their confidence levels.

We joined the rule sets from both periods and highlighted the changes in the confidence levels in both sets. We selected 18 rules found in November 2010 that also appeared in the rule set generated in November 2012. We then presented the results to the webmaster in order to find out how useful the information about the changes in the confidence would be in deciding upon possible actions to improve the website structure.

The changes in the confidence level of the association rules may reflect the changes in the behavior of the website visitors. The reasons for the confidence level change may be:

- Certain events in the institution may affect interests of the website visitors. For example, the exam periods or the time of the enrolment may affect the students' interests and the way they browse the site.
- Changes in the website structure may affect the visitor browsing patterns. This is minimized in our study since the website had undergone only minor changes.

Confidence change distribution

We categorized the rules into three categories: static, intermediate and dynamic. Rules are considered static if the confidence changed by no more than 0.03, dynamic if it changed by more than 0.1 and intermediate otherwise.

Chart 1 shows the numbers of association rules across the categories of their confidence level changes. In our sample there were 18 rules in total. Out of those, 4 rules were static, 3 were dynamic, and 11 were in the intermediate category.

Rules found in November 2010 were also categorized according to how probable it would be that a webmaster makes an action (changes the website structure) based on the knowledge presented by the rule. The rules were classified into three categories: action expected, intermediate and action not-expected.

When the webmaster was presented with the changes of the rule confidence levels based on the rules found in November 2012, we asked him/her to re-evaluate the decisions about adding/removing links on the website based on the new information. The webmaster indicated that based on the changes in the confidence level in five out of 18 rules he/she would change his/her previous decision about adding/removing links, which is shown in Chart 2.

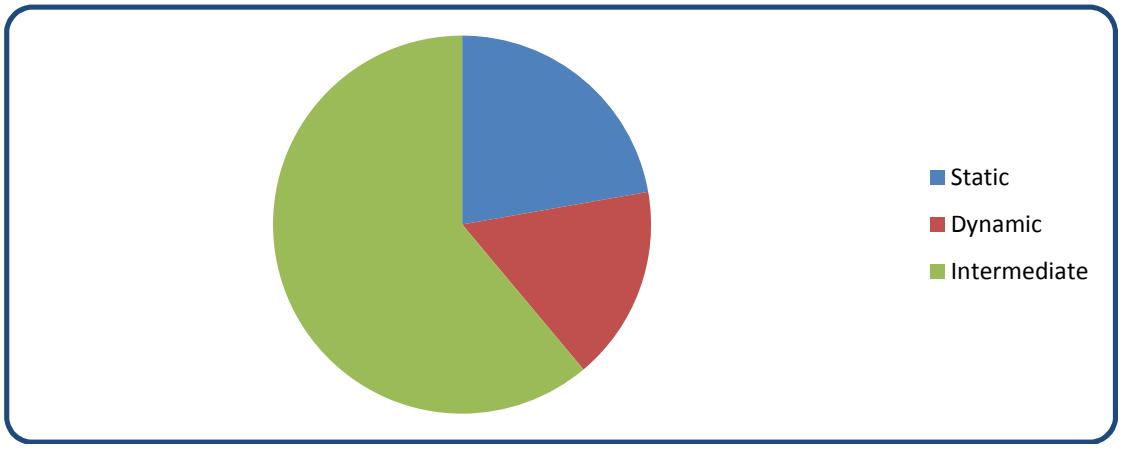

Chart 1: Dynamic vs. Static rules

The webmaster discussed that the new information about the confidence level changes was useful for all sample rules presented to him/her. It helped discriminate between the pages that are browsed in a rather consistent way and those that are browsed differently in the two time periods examined. The consistent rules confirmed the decision to create hard static links between the pages, while the changing rules pointed out pages that are more affected by the changes in the student browsing patterns.

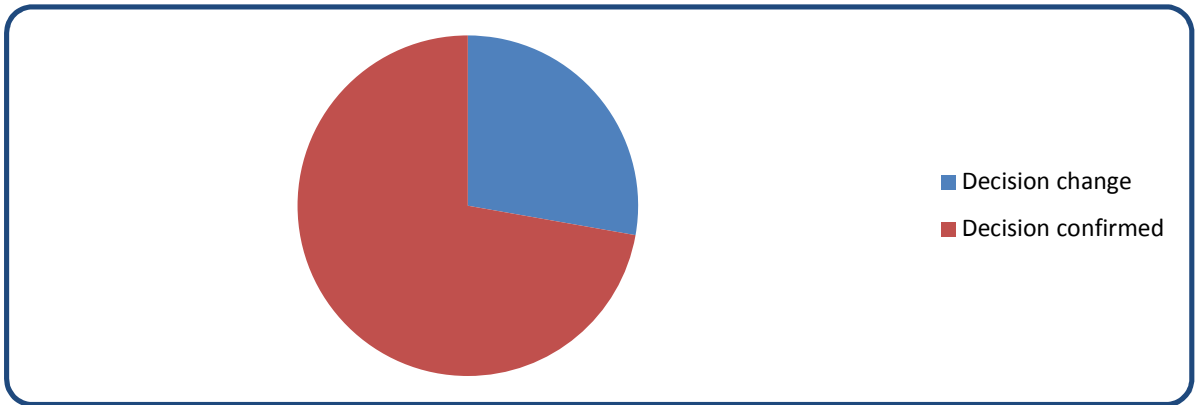

Chart 2: Rules causing webmaster's decision to change

Case discussion from the webmaster's perspective

In this paragraph, we present a sample discussion about selected rules from the webmaster's perspective.

Table 1 shows an example of two rules where the confidence levels decrease. The *Annual_exam_schedule* page contains information about the annual exam schedule, and the *Current_exam_schedule* page contains the schedule of exams in the current period. The first rule fell into the dynamic category, and the second rule fell into the intermediate category.

Page A	Page B	Confidence 2010	Confidence 2012
Annual_exam_schedule	Current_exam_schedule	0.61	0.47
Current_exam_schedule	Annual_exam_schedule	0.28	0.21

Table 1.

Based on the knowledge that students who visited the *Annual_exam_schedule* page also visited the *Current_exam_schedule* page in November 2010 with the confidence of 0.61, the webmaster had made the decision to add a new link from the *Annual_exam_schedule* page to the *Current_exam_schedule* page. However, finding that in November 2012 the confidence decreased significantly, the webmaster decided that the link should not be added. Further, the webmaster decided to watch the association between the two pages in future.

On the other hand, the opposite link from the *Current_exam_schedule* page to the *Annual_exam_schedule* page had not been added earlier, and the drop in the confidence level of the second rule confirmed the earlier decision of the webmaster.

Table 2 contains two rules between the pages related to enrollment. The *General_terms* page contains information about general terms of enrollment, and the *Undergrad_enrollment* page contains some more specific information about enrolling to undergraduate studies. The first rule fell into the static category, and the second into the intermediate category.

Page A	Page B	Confidence 2010	Confidence 2012
Undergrad_enrollment	General_terms	0.52	0.54
General_terms	Undergrad_enrollment	0.37	0.27

Table 2.

The confidence for visiting the *General_terms* page if the *Undergrad_enrollment* page is visited is high and consistent in November 2010 and November 2012. The webmaster decided that a direct link from the page *Undergrad_enrollment* to the page *General_terms* should be added.

In the opposite case the confidence for visiting the *Undergrad_enrollment* page if the *General_terms* page was visited was somewhat lower. However, since the content of two pages is related, the webmaster had considered adding a link from the *Undergrad_enrollment* page to the *General_terms page* in 2010, considering 0.37 a high enough confidence value. The drop in

the confidence in the year 2012 however, made the webmaster sure that the link would not be needed.

Conclusion

Association rule mining of web usage log files is a method that can bring new, previously unknown knowledge about the website visitor behavior. In this paper we presented how it can be used by a webmaster to improve a website structure.

We have conducted an empirical study on a website of an educational institution. Our web usage mining process consisted of four phases: pre-processing, association rule discovery, pruning, and comparing confidence levels of the rules between two time periods. The interestingness measures assigned to the association rules combined with the pruning techniques proved efficient in highlighting the rules that are easy to understand by a webmaster and prompt them to an action. The levels of changes of the rule confidence over time brought additional information to the webmaster helping them decide upon actions that may further improve the website structure.

We leave studies employing more sophisticated pruning techniques, which might exploit in more detail the interconnectedness of the web pages, as well as applying other association rule interestingness measures for future work.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, [Location?], pp. 207-216.

Anand, S. S., Mulvenna, M., & Chavielier, K. (2004). *On the deployment of web usage mining' in web mining: From web to semantic web*, pp. 23 - 42, Editors: Bettina Berendt, Andreas Hotho, Dunja Mladenic, Maarten van Someren, Myra Spiliopoulou, Gerd Stumme (3-540-23258-3), Berlin: Springer.

Balcázar J. L. (2010). Redundancy, reduction schemes, and minimum-size bases for association rules, logical methods. *Computer Science*, *6*(2:3), 1–33.

Dettmar G. (2004). *Logfile preprocessing using WUMprep*. Talk given at the Web Mining Seminar in Winter semester 2003/04, School of Business and Economics, Humboldt University Berlin, Berlin.

Dimitrijevic M., & Bosnjak Z. (2010). Discovering interesting association rules in the web log usage data. *Interdisciplinary Journal of Information, Knowledge, and Management*, *5*, 191-207.

Dimitrijevic M., & Bosnjak Z. (2011). Association rule mining system. *Interdisciplinary Journal of Information, Knowledge, and Management*, *6*, 137-150.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR), 38*(3, Article 9).

Huang, X. (2007). Comparison of interestingness measures for web usage mining: An empirical study, *International Journal of Information Technology & Decision Making (IJITDM)*, *6*(1), 15-41.

Kannan S., & Bhaskaran R. (2009). Association rule pruning based on interestingness measures with clustering. *International Journal of Computer Science Issues*, *6*(1), 35-43.

Kazienko, P. (2009). Mining indirect association rules for web recommendation. *International Journal of Applied Mathematics and Computer Science*, *19*(1), 165-186.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey, *SIGKDD Explorations, 2*(1), 1-15.

Sahaaya, A. M., & Malarvizhi M. (2010). Improving web navigation technique using weighted order representation. *International Journal of Research and Reviews in Computer Science*, *1*(2), 55-60.

Tan, P., Kumar, V., & Srivastava, J. (2004). Selecting the right interestingness measure for association patterns. *Information Systems*, *29*(4), 293 – 313.

Webb, G. I. (2011). Filtered-top- *k* association discovery. *WIREs Data Mining Knowl Discov* 2011, *1*, 183-192. doi: 10.1002/widm.28

Biographies

Maja Dimitrijević is a lecturer at the Higher Technical School of Professional Studies in Novi Sad, Serbia. She teaches courses in database structures, object-oriented programming and software engineering. She is currently working on her PhD thesis in the area of data mining. Her current research interests include data mining, web usage mining, database structures and software engineering. She holds an MSC degree in Computer Science from the University of British Columbia, Vancouver, Canada.

Tanja Krunić is a lecturer at the Higher Technical School of Professional Studies in Novi Sad, Serbia. She teaches courses in programming, web design and Internet languages and tools. She holds an MSC in mathematics and is currently working towards her PhD in Numerical Analysis from the Faculty of Mathematics and Natural Sciences, Novi Sad. Her research interests include important issues like responsive design, search engine optimization usability, accessibility, privacy, and security on the World Wide Web.