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

Volume 2, Issue 1, 2014

Improving the interestingness of web usage association rules containing common web site menu items

Maja Dimitrijevic, Higher Education Technical School of Professional Studies, Serbia, majadimi@gmail.com

Natasa Subic, Higher Education Technical School of Professional Studies, Serbia, subic@vtsns.edu.rs

Zita Bosnjak, The University of Novi Sad, Faculty of Economics Subotica, bzita.bzita@gmail.com

Abstract

The immense volume of web usage data that exists on web servers can be mined to generate association rules that contain the information about website visitor interests, which can then be utilized for enhancing the website effectiveness, or increasing the profit in e-commerce applications. One of the problems in web usage association rule mining is the generation of too many rules with extremely high statistical interestingness measures, which are not necessarily interesting to the domain expert. In this paper, we propose a modified version of added value as an interestingness measure for particular groups of association rules consisting of web pages that occur as sub-items of the web site common menus. Our real life data experiments show that the modified added value outperforms added value, resembling the domain expert defined interestingness more closely.

Keywords: web usage mining, association rules, interestingness measures, website menu

Introduction

Web server log files contain immense volumes of data, based on which the knowledge about the behavior of website visitors can be inferred, and utilized in various ways, such as enhancing the effectiveness of websites, improving the effectiveness of web marketing campaigns, or increasing the profit in e-commerce web applications. One of the popular data mining methods for automatic extraction of potentially interesting information from the web usage log files is web usage association rule mining (Kosala & Blockeel, 2000).

Association rule mining algorithms were originally applied to the analysis of transactional databases (Agrawal, Imielinski & Swami, 1993). If $I = \{i_1, \ldots, i_n\}$ is a set of items, $T = \{t_1, \ldots, t_m\}$ is a set of transactions where $t_i \subseteq I$, then an association rule is an implication of the form: $X \rightarrow Y,\ X, Y \subseteq I, X \cap Y = \emptyset$

Informally, the meaning of an association rule is "If a transaction contains a set of items X, it likely also contains a set of items Y".

The support of an association rule $X \rightarrow Y$ in the set of transactions T is defined as the probability of item sets X and Y co-occurring in the same transaction.

$Supp(X \rightarrow Y) = Count(X \cup Y)/n = P(X \cup Y)$, where $Count(A)$ is the number of transactions in T that contain the set of items A, and n is the total number of transactions in T.

The confidence of the rule R in the set of transactions T is defined as conditional probability of Y occurring in the transaction that contains X.

$$Conf(X \rightarrow Y) = \frac{Count(X \cup Y)}{Count(X)} = P(Y|X)$$

The support measures the percentage of transactions covered by the rule, while the confidence is one of the measures of how reliable an association rule is in the set of transactions.

The association rule mining algorithms generate all association rules that satisfy certain constraints, such as the minimal support constraint. Each association rule is assigned a weight, often called an interestingness measure, which specifies the degree of the potential value of the association rule for the domain expert.

In web usage association rule mining a web page of a particular website can be considered an item, while a website visitor session can be considered a transaction. The visitor session is defined as a set of web resources requested during an event of browsing (Kosala & Blockeel, 2000).

Basic steps in web usage association rule mining are data preparation, frequent set generation, and association rule generation, pruning and ranking.

Generally, association rule mining yields an overwhelming number of association rules, which makes it difficult for the data analyst to manage the association rule set, understand the rules and utilize that knowledge (Tan et al. 2004). An additional issue in web usage log association rule mining is the inherent correlation among web pages due to the website topology. Namely, the assumption about the independence between items in association rule mining is not satisfied in web usage mining due to the connectedness of web pages through hyperlinks (Huang, 2007). This causes the generation of huge numbers of association rules with very high statistical correlation, that are not truly interesting to the domain expert (Iváncsy & Vajk, 2008; Wang, Li, & Yang, 2005).

In this study, we propose an extension to the web usage mining association rule discovery process that helps the webmaster select more truly interesting rules from the set of all generated rules. The focus is particularly on association rules that contain web pages occurring as the menu items on the web site. We investigate the correlation of such web pages and propose to use a modified version of added value as the interestingness measure. Experiments are conducted on a real life data set and the performance of the proposed modified interestingness measure is evaluated by comparing it to the interestingness in terms of surprisingness and actionability assigned by the domain expert for the top N rules.

The rest of the paper is organized as follows. The following section gives an outline of the related work. The issues related to the association rules with respect to the web site topology are then discussed. The experiments with the real-life data set are then presented. The last section outlines the conclusions and directions for future work.

Related work

There are numerous statistical functions that can be used as association rule interestingness measures (Geng & Hamilton, 2006). A recent survey (Kontonasios, Spyropoulou & De Bie, 2012) discusses various approaches to interestingness, comparing subjective vs. objective and syntactical vs. probabilistic interestingness measures. Although the problem of finding the ideal

interestingness measure in various domains of association rule mining has been researched, it remains a difficult task (Jalali-Heravi & Zaïane, 2010).

Cercone and An (2002) compared the rankings of web usage association rules by 7 interestingness measures. Top 10 association rules ranked by various interestingness measures were compared to the rankings assigned by the domain expert. However, this approach suffers from generating too many rules with extremely high confidence values, which exist due to the hyperlink structure of the web site.

Jalali-Heravi and Zaïane (2010) apply 53 different statistical interestingness measures to associative classification rules, and compare them based on the number of rules and the accuracy, aiming to reduce the number of rules generated, while not jeopardizing the accuracy of the classifier.

Lee, Lo and Fu (2011) proposed to incorporate hierarchical characteristics of the website in the web usage mining process applied to the next web page prediction problem. They assume the hierarchical properties of the website are incorporated into the hierarchical folder structure of the web pages. The method decreases the size of the candidate set, increasing the efficiency of the mining process.

A line of research focuses on avoiding the minimum support threshold constraint. For example Salam and Khayal (2012) propose a method to generate top k frequent patterns.

Various techniques have been used to prune out the web usage association rule set, eliminate un-interesting rules and make it easier for the domain expert to find the most interesting rules (Dimitrijevic & Bosnjak, 2010).

Kazienko (2009) proposes mining indirect association rules in order to widen the set of web page recommendations, while not jeopardizing the recommendation accuracy. They define a variation of the confidence interestingness measure applicable to mining indirect association rules.

Association rules with respect to the web site topology

We argue that the web site topology needs to be considered in defining the web usage association rule interestingness measures, in order to alleviate the influence of the web site topology to the statistical interestingness measures. The pages that are connected by a hyperlink or through a common menu occur frequently in the same browsing sessions simply because of the way users browse the web site. That does not always reflect users' interests in the content of these web pages. Further, even when the users browse the pages that actually interest them, the association rules between hyperlinked pages or pages that are items of the same menus of the web site, do not necessarily have to be highly interesting to a domain expert (web master). Namely, such rules might not surprise the webmaster, as he/she is familiar with the web site topology and expects the correlations between those pages.

In this study, we analyze the interestingness values of correlated pages and propose a modification of an interestingness measure for a particular group of rules.

Our experiments include so called "short rules" only, where both left and right hand side of the rule contain one page only. Such rules are far easier to interpret for a domain expert, and are

often used in the literature (Kazienko, 2009). The method could be extended to the web usage association rules in general, but we leave it for future work.

The short web usage rule is denoted A→B, where A and B are web pages, as opposed to sets of web pages. According to the relation between the web pages A and B, we divide the rules into three main categories:

1) A and B are hyperlinked

Rules in this category consist of the web pages that contain a direct hyperlink from one to another. Such rules usually have the highest statistical correlations, with confidence often close to 100%. In our experiments we prune out the rules in this category. There are methods that can be used to assess the usefulness of hyperlinks based on such rules (Kazienko & Pilarczyk, 2006), which we leave out of the scope of this paper.

2) A and B are items of the same sub-menu

The focus of our study is the rules between web pages that occur as sub-items of the same web site menu item. Such rules usually have high statistical correlations, but not close to 100%. The sub-items of the same menu item are often browsed together, which can be the result of user interests in the content of those pages, but can also be the result of the way the menu is organized. Further, even when the users are really interested in the content of those pages, the domain expert (webmaster) will often find such rules not highly interesting, as he/she usually intentionally organizes the menu items so that the sub-items are content related. Therefore, an association rule might only confirm the content relatedness of the sub-items, and does not surprise the webmaster. However, the rules in this category, that have extreme values (both positive and negative) of statistical interestingness measures may be of interest to the webmaster.

We limit our experiments to the rules between the pages that are on the same level of the menu, not considering deeper levels of the menu hierarchy. The rules between such pages and the meaning of their statistical interestingness measures are relatively easy to interpret by the domain expert. We plan to include analyzing deeper levels of menu hierarchy in future work.

3) A and B are not tightly linked

The third category contains the rules where there is no direct hyperlink from one page to the other, and A and B are not sub-items of the same menu item. These pages do not have high expected statistical correlations and common association rule interestingness measures can be applied. Therefore, we did not include rules in this category in our experiments.

Experiments

We conducted an experiment on a real life data set containing the actual visits to the web site of the Higher Education Technical School of Professional Studies at www.vtsns.edu.rs in the month of December, 2013. The web log file contained 1,367,603 web resource requests.

We cleaned the web usage log file by removing irrelevant and robot requests, which resulted in 68,660 requests. The requests were divided into browsing sessions, where a session is defined as a set of requests coming from the same IP, while the time between two requests does not exceed the threshold of 5 minutes. This resulted in 24,771 browsing sessions.

Association rule generation

We generated association rules using the software (Dimitrijevic & Bosnjak, 2011), which implements a version of the well-known Apriori algorithm for frequent set generation. The support threshold was set to 0.001 and the confidence threshold to 0.1. We chose very low parameter values in order to minimize their influence on the rules discovered. Our goal was to find as many rules as possible, even the non-frequent ones, and focus on analyzing the influence of the menu structure to the interestingness measure values.

This resulted in 874 association rules in total. We then extracted only the rules that contain web pages appearing on the same sub-menu of the main common web site menu. The number of such rules was 127.

To evaluate the interestingness of association rules we used added value (AV) as a statistical interestingness measure. It is one of the commonly applied interestingness measures, that is easy to understand by the domain expert who evaluates the generated association rules. Added value slightly differs from confidence in that it shows the difference of the reliability of the association rule and the coverage of the right hand side across all transactions. Basically, it shows the difference of the probability that Y occurs in the transactions that contain X, and the overall probability of Y occurring in the set of all transactions. It is given by the following formula.

$$AV(X \rightarrow Y) = \frac{Count(X \cup Y)}{Count(X)} - \frac{Count(Y)}{n}$$

$$AV(X \rightarrow Y) = Conf(X) - Supp(Y)$$

Domain expert defined interestingness

The real interestingness of generated association rules is often evaluated by users, that is by the domain experts (Geng & Hamilton, 2006). In this study, we define the association rule interestingness for the domain expert (webmaster) with respect to two notions: *surprisingness* and *actionability*. We consider a rule R surprising for the webmaster if it draws his/her attention to the user behavior that cannot be explained by the website topology. We consider a rule R actionable if it helps the webmaster make a decision about changing the website structure (adding/deleting links or re-organizing the menus).

In most cases the rule interestingness coincides with the rule actionability. However, sometimes the rule brings the webmaster's attention to an unexpected behavior of users, which does not necessarily initiate the webmaster to make an action on the website.

It is usually rather difficult for a domain expert to exactly rank top N association rules by their interestingness. Often several rules are almost equally interesting to the domain expert. Instead, we asked the webmaster to assign a value of interestingness (i.e. surprisingness and actionability) to each of the top 15 rules when the rules. In order to make the task easier for the webmaster, we limited the surprisingness and actionability to the set of values {0, 1, 2, 3}, with 0 meaning "not surprising/actionable", 1 meaning "somewhat surprising/actionable", 2 meaning "surprising/actionable", and 3 meaning "extremely surprising/actionable". We define the overall

expert interestingness (EI) of a rule R as the average of assigned surprisingness and actionability values.

$$EI(R) = (Surp(R) + Act(R))/2$$

It would be possible to assign the weights of the surprisingness and actionability in the expert interestingness formula, but it was not the focus of our experiment, and we decided to use the simple average formula.

Categories of association rules occurring as menu sub-items

If a short association rule is of the form A→B, where A and B are sub-items of the same menu item of a common menu on the web site, we consider 4 categories based on the relative position of pages A and B on the menu:

1) Category A-B

A and B are one next to each other on the menu, and A appears before B. If the user browsed the menu in a common way without skipping the items, he/she would click the item A, and then the item B.

2) Category B-A

A and B are one next to each other on the menu, and B appears before A. If the user browsed the menu in a common way without skipping the items, he/she would click the item B, and then the item A.

3) Category A...B

There are one or more items between A and B on the menu, and A appears before B. If the user browsed the menu in a common way without skipping the items, he/she would click the item A, then some other items, and then the item B.

4) Category B...A

There are one or more items between A and B on the menu, and B appears before A. If the user browsed the menu in a common way without skipping the items, he/she would click the item B, then some other items, and then the item A.

The average added value for the set of rules in each of the four categories is shown in Table 1.

	A-B	**B-A**	**A...B**	**B...A**
Average AV	0.256	0.401	0.109	0.251

Table 1: Average added value for each category of rules

The average added value for all rule categories are far beyond 0, which would be expected if there was no statistical correlation between pages A and B. The average for the category B-A is much higher than for the other categories. This can be explained by the natural way users browse the menu. If the menu item A appears right after the item B on the menu, most visitors who click on the item A would have also clicked the item B while browsing the menu items in the order

they appear on the menu. Further, the web master often places content related web pages close to each other on the menu, so they can be co-occurring in the same browsing sessions due to their content relatedness as well.

On the other hand, the average for the category A-B is somewhat lower, even though the items are next to each other on the same menu. This can also partly be explained by the pattern of browsing. Many visitors who click the item A, which occurs before the item B on the menu stop browsing the menu, and never get to the item B.

As expected, the averages for the categories where there are other items between A and B on the menu are somewhat lower. Similarly to the first two categories, the rule category where B appears before A has higher average added value than the one where A appears before B.

The proposed modification to statistical interestingness

We propose a modified version of added value, which we refer to as MAV. The idea is to penalize the statistical interestingness values for the rules in the categories with high average added value. We consider the average as an approximation of an expected value for the rules in the category. The webmaster is more likely to be interested in the rules with extreme values of added value, as opposed to the rules with values closer to average.

A simple method is to subtract the average AV for the rule category from the actual AV for the rule. We assume that the rules with added value close to the average for the category are not very interesting to the web master.

We define the modified added value of a rule as follows:

$$MAV(R) = AV(R) - Avg(CR), where\ R \in CR, and\ CR\ is\ a\ rule\ category$$

Evaluating modified added value

It is assumed that the more similar a statistical interestingness measure is to the expert defined interestingness, the better it is to use in the particular domain.

We compare the expert interestingness of top N association rules ranked by MAV with the expert interestingness of top N association rules ranked by AV.

Table 2 shows the number of rules among the top 15 rules that fall into each of the 4 categories of surprisingness (Surp) and actionability (Act), when the rules are ranked by added value (AV), Modified Added Value (MAV) and Expert Interestingness (EI).

	AV		MAV		EI	
Category	**Surp**	**Act**	**Surp**	**Act**	**Surp**	**Act**
3-Extreme	1	1	2	2	2	2
2-High	6	5	7	6	8	6
1-Low	8	8	6	6	5	7
0-Zero	0	1	0	1	0	0

Table 2: Distribution of surprisingness and actionability of the top 15 rules ranked by the three interestingness measures

As shown in Table 2, the distributions of the top 15 rules over surprisingness and actionability categories are similar for all three interestingness measures. Added value seems just slightly worse than the modified added value, as in the top 15 according to added value at least 2 highly surprising or actionable rules are missing.

However, the exact rankings of the rules among the top 15 in our experiment vary for the three interestingness measures. In order to evaluate how well MAV ranks the top N rules, we consider the vectors of EI values (average of surprisingness and actionability) for top N rules ranked by AV, MAV, and EI.

Let M^N be the vector of EI values for top N association rules ranked by an interestingness measure M. The *i-th* element of the vector is the value of EI for the *i-th* rule ranked by the interestingness measure M.

$$M_i^N = EI(R_i)$$

To evaluate the distance of rankings between two interestingness measures M1 and M2, we can measure the distance between the corresponding vectors $M1^N$ and $M2^N$ based on the Euclidean distance formula as follows:

$$Distance(M1^N, M2^N) = \sqrt{\sum_{i=1}^{N}(M1_i^N - M2_i^N)^2}$$

Table 3 shows the distance of ranking vectors by MAV and EI, and the ranking vectors by AV and EI for top 5, top 10 and top 15 rules.

	Top-5	**Top-10**	**Top-15**
Dist(MAV,EI)	1.58	2.00	2.24
Dist(AV,EI)	2.55	2.87	3.00

Table 3: The distance between ranking vectors

The distance of MAV to EI is lower than the distance of AV to EI for each of the Top-N rules in this experiment. This result supports the hypothesis that the top N rules ranked by MAV are more interesting to the domain expert than the top N rules ranked by AV.

Conclusion

In this study, we analyzed the correlation of web pages due to the web site topology. We argued that the web site topology needs to be considered in defining the web usage association rule interestingness measures, since it heavily influences statistical interestingness measure values. The pages that are connected by a hyperlink or are the sub-items of the same menu item are visited frequently in the same browsing sessions simply because of the way users browse the web site. That does not necessarily correspond to the association rule interestingness defined by the domain expert.

We defined the association rule interestingness for the domain expert (webmaster) with respect to two notions: *surprisingness* and *actionability* and compared it to the statistical interestingness measures.

We proposed a modified version of added value as an interestingness measure for rules containing web pages that occur as sub-items of the same menu item on the web site. We conducted experiments on a real life data set to evaluate how well the modified added value (MAV) ranks the rules, comparing to the rankings by the common added value (AV). The experiments confirmed that the interestingness of the top N rules ranked by MAV is closer to the domain expert defined interestingness, compared to the interestingness of the top N rules ranked by AV.

Our experiments are conducted on so called "short rules" only, where both left and right hand side of the rule contain one page. The method could be extended to the web usage association rules in general, which we leave for future work.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.

Becker, K., & Vanzin, M. (2010). O3R: Ontology-based mechanism for a human-centered environment targeted at the analysis of navigation patterns. *Knowledge-Based Systems*, *23*(5), 455-470.

Cercone, N., & An, A. (2002, November). Comparison of interestingness functions for learning web usage patterns. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 617-620). ACM.

Dimitrijevic M., Bosnjak Z. (2010). Discovering Interesting Association Rules in the Web Log Usage Data, Interdisciplinary Journal of Information, Knowledge, and Management, Volume 5, pages 191-207. http://www.ijikm.org/Volume5/IJIKMv5p191-207Dimitrijevic443.pdf

Dimitrijevic M., Bosnjak Z. (2011). *Association Rule Mining System*, Interdisciplinary Journal of Information, Knowledge, and Management, Volume 6, pages 137-150. http://www.ijikm.org/Volume6/IJIKMv6p137-150Dimitrijevic552.pdf.

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

Huang, Xiangji (2007). Comparison of Interestingness Measures for Web Usage Mining: An Empirical Study, International Journal of Information Technology & Decision Making (IJITDM), vol. 06, issue 01, pages 15-41.

Jalali-Heravi, M., & Zaïane, O. R. (2010, March). A study on interestingness measures for associative classifiers. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (pp. 1039-1046). ACM.

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

Kazienko, Przemysław, and Marcin Pilarczyk. (2006). "Hyperlink assessment based on web usage mining." *Proceedings of the seventeenth conference on Hypertext and hypermedia*. ACM, 2006.

Kontonasios, K. N., Spyropoulou, E., & De Bie, T. (2012). Knowledge discovery interestingness measures based on unexpectedness. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(5), 386-399.

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

Lee, C. H., Lo, Y. L., & Fu, Y. H. (2011). A novel prediction model based on hierarchical characteristic of website. *Expert Systems with Applications*, *38*(4), 3422-3430.

Merceron, A., & Yacef, K. (2008). Interestingness Measures for Associations Rules in Educational Data. *EDM*, *8*, 57-66.

Salam, A., & Khayal, M. S. H. (2012). Mining top− k frequent patterns without minimum support threshold. *Knowledge and information systems*, *30*(1), 57-86.

Senkul, P., & Salin, S. (2012). Improving pattern quality in web usage mining by using semantic information. *Knowledge and information systems*, *30*(3), 527-541.

Tan, P., Kumar, V., Srivastava, J. (2004). Selecting the Right Interestingness Measure for Association Patterns. Information Systems, Volume 29, Issue 4, Pages: 293 – 313.

Biographies

Maja Dimitrijevic 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.

Natasa Subic is a professional associate at the Higher Technical School of Professional Studies in Novi Sad, Serbia. She teaches courses in Introduction to Programming, Internet languages and tools, Computer Graphics and Computer Animation. Her current research interests include Web Responsive Design, Apply animation in professional education and Modeling of flexible and adaptive business systems. She holds an M.Sc. degree in Computer Science from the University

of Novi Sad, Serbia and degree for Specialist Engineer of Organizational Sciences, University of Belgrade, Faculty of organizational sciences, Serbia.

Zita Bosnjak is a full professor at the University of Novi Sad, Faculty of Economics Subotica, Department of Business Information Systems and Quantitative Methods. She received a B.S. (1987) in Informatics from the University of Novi Sad, Faculty of Sciences and an M.S. (1991) and a Ph.D. (1995) in Informatics from the University of Novi Sad, Faculty of Economics Subotica. Her current research interests include the theory and practice of knowledge in data discovery and expert and fuzzy systems, and their application to business, strategic management, education and capacity building. She has written over 20 journal articles, 3 books, and 50 conference articles on related topics. From January 2006 she has been a member of the editorial board of the Management Information Systems journal.