Discovering Interesting Association Rules in the Web Log Usage Data

Maja Dimitrijević
The Advanced Technical School, Novi Sad, Serbia

dimitrijevic@vtsns.edu.rs

Zita Bošnjak
The University of Novi Sad, Faculty of Economics
Subotica, Serbia

bzita@eccf.su.ac.yu

Abstract

The immense volume of web usage data that exists on web servers contains potentially valuable information about the behavior of website visitors. This information can be exploited in various ways, such as enhancing the effectiveness of websites or developing directed web marketing campaigns. In this paper we will focus on applying association rules as a data mining technique to extract potentially useful knowledge from web usage data.

We conducted a comprehensive analysis of web usage association rules found on a website of an educational institution. Our experiments confirm that, prior to pruning, the set of generated association rules contained too many non-interesting rules, which made it very difficult for a user to find and exploit useful information. Many of these rules are a simple consequence of the high correlation between web pages due to their interconnectedness through the website link structure.

We proposed and applied a set of basic pruning schemes to reduce the rule set size and to remove a significant number of non-interesting rules. This pruning method decreased the size of our experimental rule set by more than three times, making it much simpler to browse for truly interesting rules. The percentage of truly interesting rules, which can initiate a webmaster to actions that can potentially enhance the website and improve its browsing experience, in our resulting experimental rule set was 41%.

The analysis of association rules in our case study confirmed the hypothesis that discovering interesting and potentially useful association rules in web usage data does not have to be a time-consuming task and can lead to actions that increase the website's effectiveness.

Keywords: association rules, web usage data, pruning, interestingness measures, website link structure

Introduction

Due to the immense volume of Internet usage and web browsing in recent years, log files generated by web servers contain enormous amounts of web usage data that is potentially valuable for understanding the behaviour of website visitors. This knowledge can be applied in various ways, such as enhancing the

Editor: Eli Cohen

effectiveness of websites through user personalization or developing directed web marketing campaigns (Anand, Mulvenna & Chavielier, 2004; Cooley, Mobasher, & Srivastava, 1997).

Data mining methods, which, by definition, are suitable for automatic extraction of potentially interesting information from very large databases, are used to extract knowledge from the web usage log files. One of the popular data mining methods that has been used for this purpose is association rule finding (Kosala & Blockeel, 2000).

Originally, association rule mining algorithms were applied for the analysis of transactional databases (Agrawal, Imielinski, & Swami, 1993).

An association rule is defined as follows:

Let $I = \{i_1, ..., i_n\}$ be a set of items, and $T = \{t_1, ..., t_m\}$ a set of transactions, where each transaction t_i consists of a subset of items in I. An association rule is then an implication of the form:

$$X \to Y, X \in I, Y \in I, X \cap Y = \varnothing$$

An item set X has support s in T if $s\%$ of the transactions in T contains X.

An item set X is frequent if its support is higher than the user specified minimum support.

The rule $X \to Y$ holds in T with confidence c if $c\%$ of transactions in T that contain X also contain Y.

The problem of mining association rules is to generate all association rules that consist of frequent item sets and the confidence greater than the user-specified minimum confidence.

While association rule finding algorithms are complete in that they find all rules that satisfy defined constraints, they often result in a large set of rules that is difficult to exploit and find those rules that are truly interesting to the user. Various methods have been proposed to help deal with this issue.

For example, a query language called “Mine Rule”, originally developed for querying inductive databases, can be applied to mining the set of generated association rules (Meo, Luca Lanzi, Matera, Careggi, & Esposito 2004). Furthermore, various methods have been proposed to prune the set of generated rules and discard irrelevant rules (Jaroszewicz & Simovici, 2002; Liu, Hsu, & Ma, 1999). Another area of research focuses on finding various association rule ‘interestingness measures’, which help find the rules that give maximally useful information to the user in the set of generated association rules (Tan, Kumar, & Srivastava, 2004). Some of the proposed association rule interestingness measures are a*ll-confidence* (Omiecinski, 2003), *collective strength* (Aggarwal & Yu, 1998), *conviction* and *lift* (Brin, Motwani, Ullman, & Tsur, 1997).

When applying association rule mining to web usage data, a web resource of a particular website is usually considered an item, while a website visitor session is considered a transaction of items. Here, a website visitor session is a set of web resources that a visitor requested during one event of browsing the website (Anand et al., 2004).

Although various interestingness measures and rule pruning methods have been applied to association rule mining of web usage data, extracting useful information from the set of generated association rules remains a difficult task (Geng & Hamilton, 2006; Huang, 2007).

Web usage data is specific and differs from the market basket data 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 occur in the same transaction, which 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. Some researchers have abandoned the association rule model and have chosen instead to concentrate on sequential patterns or other data mining techniques to mine web usage data (Iváncsy & Vajk, 2008; Wang, Li, & Yang, 2005).

In this paper, we investigate how effective association rules can be when discovering potentially useful information in the web log data of an educational institution. We were especially interested in the information that can prompt actions leading to enhancing a website and improving the browsing experience for visitors. In relation to this, another goal of our study was to investigate what needs to be done with a set of association rules generated by the rule-finding algorithm, so that extracting useful knowledge from the web log data becomes a task worth carrying out.

The information discovered by association rule mining of web log usage data can potentially be used by webmasters when enhancing the effectiveness of the websites. Through our experiments, we were able to demonstrate how association rules found while mining the web usage data of an education institution can be used by a webmaster to increase the usability of the institution's website.

In our case study, we used raw web log data of the Advanced Technical School in Novi Sad. We first prepared data for association rule finding using various web log data cleaning and transformation tools (Detmar, 2004). We then ran an association rule finding algorithm implemented in Weka, one of the most popular open source data mining tools (Weka, n.d.). Finally, we proposed and applied a method for basic pruning of trivial association rules on the generated association rules of the institution's web usage data. Having the knowledge of the structure of the institution's website and the behavior of the site's visitors, we then analyzed the resulting association rule set from the user's point of view and considered how well the interestingness measures coincided with the subjective interestingness of the found rules before and after the pruning.

The rest of the paper is organized as follows: We will discuss various phases of association rule mining of the Web usage data, while at the same time presenting the results of this process on our experimental data set in the section *Experiment 1: Finding Web Usage Association Rules*. The section *Schemes for Pruning Association Rules* will define the schemes that we proposed and applied for pruning the rule set generated in our experiment. We will then present the effects of this pruning on our rule set in the section *Experiment 2: Pruning Web Usage Association Rules*. Next, we will discuss the value of the rule set from a webmaster's point of view in the section *Interestingness of the Resulting Association Rules*. Finally, the *Conclusion* will give an overview of the results.

Experiment 1: Finding Web Usage Association Rules

In this section, we will discuss the two phases of Web usage association rule mining – data preparation and applying association rule algorithm – while presenting the effect of this process on our experimental real life data set.

In our case study, we found association rules in the web usage data of the Advanced Technical School in Novi Sad. Since most visitors of this website are students, we expected to find rules that correlate to web pages that contain information about exam schedules, grades, announcements etc.

The purpose of this experiment was to give some insight into the usefulness of association rules when they are applied to the web log data set of an education institution. As a part of the experiment, our goal was to compare the association rule set before and after the pruning and thus to evaluate basic pruning techniques that we propose for this particular kind of data.

Data Set

For experimental purposes, we used the log file containing information about all web requests to the institution's official website on November 16, 2009. This was an arbitrarily selected day, and we were not aware of any special activities at the institution on that day which could have led to any unusual behaviour of the website visitors.

Each line in the web usage log file contains information about one web resource request, the time of request, the URL requested, as well as other information (IP, web browser info, etc.) that can be ignored when mining association rules of the web pages requested in various sessions. The raw web log file we used for the experiment contained 5999 web requests. This file can be found at http://www.vtsns.edu.rs/maja/vtsnsNov16 .

After the data preparation process, the file contained 426 user sessions (sets of pages visited during the same visit to the website).

Data Preparation

In order to prepare the web log data for the mining process, the web log file needed to be cleared of irrelevant requests, each relevant request needed to be assigned to a visit session, and the resulting file had to be transformed to a format that could be fed into the mining algorithm.

For cleaning the web log file of irrelevant requests and creating sessions in it, we used a freely available tool for web log data preparation called WumPrep, which consists of a set of Perl scripts (Dettmar, 2004).

Removing irrelevant requests

The first step in preparing a log file for data mining is removing irrelevant requests such as images, icons, and other resources that are embedded in a web page. For the purpose of finding rules or patterns in web usage data, we are only interested in the pages or documents the visitors visit when traversing a website.

In order to remove irrelevant resources from our web log file, we used a WumPrep's Perl scripts, which removed 3772 irrelevant requests from the log file, out of the original 5999, leaving 2227 web page or document requests.

Removing automatic requests

Various software robots, indexers, and spiders make automated requests while crawling the World Wide Web and creating their own databases. These requests are logged in the web log file but are not representative of the behavior of actual visitors to the website and would make noise in the analysis process.

To remove automatic requests we used a WumPrep's Perl script, which recognizes robot requests based on certain heuristics (Dettmar, 2004). However, since the script was somewhat outdated and would not be able to recognize all robot requests, we did some additional manual cleaning of obvious robot requests. Cleaning of the automatic requests resulted in a file containing 2122 requests (out of 2227) made by actual visitors. We removed 105 out of 2122 requests, which is 4.9%. It is possible that we have removed a few relevant requests, which we estimate at less than 1.0% of the original data set size. However, we did not break any visitor sessions by removing requests from them during this process. Instead, we removed complete visitor sessions only. Therefore this process may have somewhat decreased the data set size (by less than 1% in our case), but it did not influence the validity of the results.

There are scripts available for purchase that can accurately remove all automatic requests, which is something that ought to be done before conducting an experiment on a larger data set. However, for the purpose of our experiments and testing the validity and effectiveness of association rule mining and pruning based on the proposed schemes, the amount of manual work was manageable.

Creating sessions

The term session in Web usage mining assumes a series of web requests made by a user during her/his visits to a website. Several methods were used for identifying the beginning and end of visitor session. WumPrep's script, which we used to define sessions, defines a single session according to a user specified maximal page view time (Dettmar, 2004). After applying this tool, the 2122 requests in our web log data file were organized into 426 sessions.

Transforming the file format

The association rule finding algorithm in (Weka3) accepts input files in a format called Arff (Weka, n.d.). An Arff file has a header containing the list of all attributes and a list of transactions, each of which contains the list of all attributes and their values in each transaction. This more naturally corresponds to mining data in relational tables then the market basket type data, which is analogous to web session data in our case.

There are two possible representatives of data in the Arff file – dense and sparse. In both formats a web page must be considered as a binary attribute that takes the values true or false in each transaction, depending on whether the page occurs in the transaction or not. Since the occurrence of web pages in transactions (user sessions) is scarce, we found the sparse format to be more appropriate for presenting sessions in web log data.

Since WumPrep (or any other tool that we could find) does not contain any script that converts web log data into either sparse or dense Arff file format, we developed a Perl script for this purpose.

The Arff file that resulted from our data preparation that we used in our experiments can be found at http://www.vtsns.edu.rs/maja/vtsnsNov16.arff .

Applying Association Rule Algorithm

Association rule mining tool

In order to conduct our experiments, we used Weka's implementation of the Apriori association rule finding algorithm based on Agrawal and Srikant (1994). In addition to setting minimum support for each item set and the minimum interestingness of a rule, the algorithm allowed us to set the maximum number of generated rules.

The algorithm supports 4 different types of interestingness measures: confidence, lift, leverage, and conviction. While lift and leverage are symmetric in regards to the left and right hand side of the rule, confidence and conviction have different values for the rules of the form X==>Y and Y==>X. We found that the symmetric measures do not give us potentially useful information about the association rules in the context of web usage mining domain. Therefore, we used confidence and conviction when analyzing the interestingness of the found association rules.

We used KnowledgeFlow interface in Weka, which allows for assembling of various types of modules that are a part of the data mining process. We loaded a file in Arff format, viewed it using a text viewer module for validating purposes, ran it through the association rule finding module, out of which it flowed to the text viewer to show the results, as shown in Figure 1.

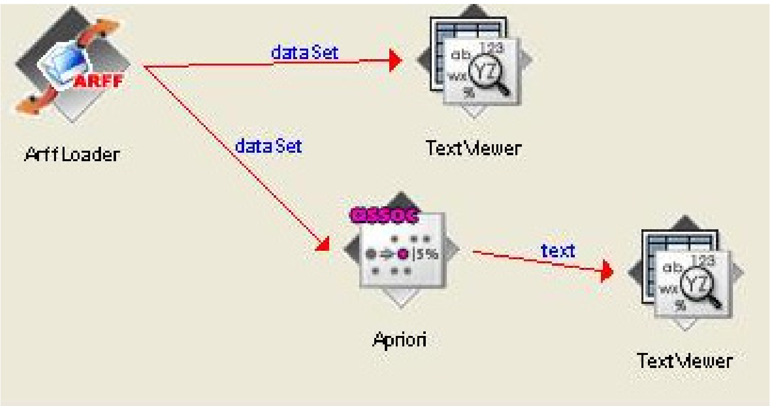

Figure 1: Data flow in Weka

Setting parameter values

While conducting the experiments, we noticed that a lot of interesting rules contained item sets with support of less than 0.1, which is a default value in Weka. Based on our empirical research, we chose to set the minimum support of an item set to 0.08. Out of the total of 34 items (web resources that occurred in our transactions after the log file preparation) our data set contained 12 frequent items according to the chosen minimum support. The maximum size of a frequent item set was 5 items.

The algorithm calculated the values of all four interestingness measures available in Weka. We chose to set the minimum value for conviction of an association rule to 1.1 (Brin et al., 1997). At the same time, we set the maximum number of generated association rules to be 300. The minimum value of confidence in any rule generated by the algorithm was 0.17.

The generated rule set

In accordance with our expectations, the initially generated association rule set contained many rules that had very high confidence. There were 19 (out of 300) rules with confidence equal to 1.0, while 69 (out of 300) rules had confidence greater than 0.85. This can be explained by the fact that many web pages are strongly correlated due to the link structure of the website.

Challenges in understanding the rules

Considering the meaning of the web pages and their connectedness through the link structure of the School's website, we browsed the set of the 300 generated rules in order to identify those that are truly interesting, i.e., those that give potentially useful information about the behaviour of the website visitors. Two main obstacles we faced were as follows.

1. The number of potentially interesting rules generated by the Apriori algorithm that we would not identify as truly interesting was overwhelming. Browsing through so many rules that seemed not to give any useful information, it was quite difficult to find rules that could potentially be truly interesting.
2. The interestingness measures were not reliable enough to reflect the true interestingness of the rules.

The next section describes the method we chose to tackle the first obstacle in understanding the rules and finding the truly interesting ones.

Schemes for Pruning Irrelevant Rules

We found out that most of the rules that were not truly interesting were actually related to "strong" rules, which have very high confidence, such as this:

/CurrSch.doc ==> /CurrSch.php , conf: 0.99, lift: 5.25, lev: 0.13, conv: 28.02

(CurrSch is an abbreviation of Current Exam Schedule)

Unlike classic market basket, there are many "strong" rules in web usage data due to the strong connectedness of the web pages (items in the rules) through the link structure of the website.

Therefore, the first step needed before looking at the interestingness measures of the rules was to prune out the trivial or irrelevant rules, in order for browsing for relevant and interesting rules to become possible.

In the following section, we will give a formal description of simple schemes that we proposed to use for pruning trivial or irrelevant rules generated by Apriori algorithm in Weka. The schemes are simple and straightforward and can be used for developing a software tool that could be applied to prune the association rule set generated by Weka. For the purpose of our experiment, we manually applied the schemes to prune the rules generated on our sample data set in Weka, in order to test their validity and effectiveness.

Rule Pruning Schemes Based on Page Clusters

We defined the clusters of web pages according to the generated association rules that have very high confidence (close to 1). Unlike association rule mining in the domain of classic market basket data, in the domain of web usage data the clusters of pages exist due to "hard" connectedness of some pages through the link structure of the website.

For example, if there is a link from page *a* to page *b*, it is likely that many visitors of page *a* will also visit page *b*. On the other hand, if page *a* is the only page that has an incoming link to page *b*, most visitors will have to visit page *a* in order to get to page *b*. Therefore, the set of generated association rules will contain both rules *a==>b* and *b==>a*, which may both have confidence close to 1.

Definition 1: Page cluster

Let us suppose that the set of all rules R contains the following rules:

a==>b, conf(a==>b) ≈ 1

b==>a, conf(a==>b) ≈1 , where a and b are items $a \in I, b \in I$

We define a cluster C^{ab} = {*a*, *b*}

The Rule Pruning Schemes

Let a and b be items, and C^{ab} their cluster according to Definition 1.

We will use the notation "a X" for a set of items "{a} ∪ X", where X is any item set.

We define the following schemes that relate to the rules containing items a and b, which belong to a cluster C^{ab}.

Scheme 1

Let Y be a non-empty item set, and X any item set.

Let us consider the following five rules in the rule set R:

1. a X ==> b Y
2. b X ==> a Y
3. a X ==> Y
4. b X ==> Y
5. a b X ==> Y

Note: The notation "a X==>Y" and "X==>b Y" assumes any rule that contains item a on the left hand side, i.e., item b on the right hand side respectively, since the items are not ordered either on the left or right hand side of an association rule.

If the above rules exist in the rule set R and have approximately the same confidence, we consider them equivalent and redundant. They can be pruned out and exchanged by using the following rule representative without losing any information in the rule set R:

Scheme1 rule representative: C^{ab} X ==> Y

Since the five rules that fit into Scheme 1 have similar but not equal confidence, we choose to assign a median value of confidence of the five rules to the cluster rule representative. Example rules that satisfy Scheme 1, their cluster representatives, and their confidence are provided in the section *Experiment 2: Pruning Web Usage Association Rules* and can also be found at http://www.vtsns.edu.rs/maja/.

Rationale

Let A^T be a set of transactions that contains the item a, and B^T a set of transactions that contains the item b.

If X and Y are sets of items that are impartially distributed over the sets A^T and B^T, the confidence of all rules defined by Scheme 1 is approximately the same. The differences in the confidences of the rules in Scheme 1 are then small and exist only due to the fact that the rules a==>b and b==>a have confidence somewhat smaller than 1.

If there is a significant difference in the confidences of the rules in Scheme 1, the rules of the Scheme 1 may be interesting to the user and, therefore, cannot be pruned out.

Scheme 1a

Let us consider the following two rules in the rule set R:

1. a X ==> b
2. b X ==> a

Note: Scheme 1a is a special case of Scheme 1, where Y is an empty set.

If the confidence of the above rules is close to 1, they are simply a consequence of the rules a==>b, conf=1 and b==>a, conf=1 and can, therefore, be pruned out of the rule set as being non-interesting.

Scheme 2

Let X be a non-empty item set, and Y any item set.

Let us consider the following three rules in the rule set R:

1. X ==> a Y
2. X ==> b Y
3. X ==> a b Y

If the above rules exist in the rule set R and have approximately the same confidence, we consider them equivalent and redundant. They can be pruned out and exchanged by the following rule representative without losing any information in the rule set R.

Scheme 2 rule representative: X ==> C^{ab} Y

The rationale behind this pruning is analogous to the Rationale behind the pruning according to Scheme 1.

Since the three rules that fit into Scheme 2 have similar but not equal confidence, we chose to assign median value of confidence of the three rules to the cluster rule representative. Example rules that satisfy Scheme 2, their cluster representatives, and their confidence are given in the section *Experiment 2: Pruning Web Usage Association Rules*. The list of the resulting rules before and after the pruning process, as well as their cluster representatives can be found at http://www.vtsns.edu.rs/maja/. Note that we renamed the web pages in order to increase readability of the rules, while not changing the structure of the data.

Experiment 2: Pruning Web Usage Association Rules

Removing Rules that Contain Home Page

As the first pruning step, we removed all rules that contain the website home page. The home page itself has support close to 1.0 (contained in almost all user sessions) and the set of all transactions is almost equivalent to the set of all transactions that contain the home page. Therefore, the rules containing the home page are a consequence of their sub-rules. After removing the rules that contain the home page, our rule set contained 96 rules out of the 300 rules generated originally.

Identifying Page Clusters

In our association rule set previously described in the section *Applying Association Rule Algorithm*, we identified two rules that define a two-page cluster according to Definition 1. The rules are as follows:

/CurrSch.doc ==> /CurrSch.php , conf: 0.99

/CurrSch.php ==> /CurrSch.doc , conf: 0.85

We define a cluster that we call CurrSch to represent the pages { CurrSch.php , CurrSch.doc }.

It is worth noting that when looking at the structure of our website we find that the above rules exist in our data set because the page CurrSch.php has only one outgoing link, which is the link to the page CurrSch.doc. At the same time, the page CurrSch.php is the only incoming link to the page CurrSch.doc.

Pruning According to Scheme 1a: Eliminating Trivial Rules

Considering the cluster CurrSch, we identified 10 rules that can be eliminated as being trivial according to Scheme 1a. The confidence of all eliminated rules was in the range of 0.86 to 1.0.

An example of such a rule is:

/CurrSch.doc /AnnualSch.php /exams.php ==> /CurrSch.php , conf: 1

Pruning According to Scheme 1: Exchanging Sets of Rules by Their Cluster Representative

When pruning the rules in our sample data set we identified 8 clusters of rules that fit into Scheme 1. We eliminated 40 rules and introduced 8 rules as their cluster presentation (1 for each cluster), thus decreasing the size of the rule set by 32 rules (out of 96).

For example, we eliminated the following five rules and introduced their cluster representatives. The confidence of all eliminated rules is close, which means they all fit into the Scheme 1.

/CurrSch.doc /ads.php ==> /CurrSch.php /exams.php , conf: 0.98

/CurrSch.doc /ads.php ==> /exams.php , conf: 0.98

/CurrSch.doc /CurrSch.php /ads.php ==> /exams.php , conf: 0.98

/CurrSch.php /ads.php ==> /exams.php 50 , conf: 0.98

/CurrSch.php /ads.php ==> /CurrSch.doc /exams.php , conf: 0.84

The cluster representative of the five eliminated rules is:

CurrSch /ads.php ==> exams.php , conf: 0.91

In our rule set, all rules that followed the format of Scheme 1 had similar confidence. Therefore, we were able to eliminate five non-interesting rules for every rule-cluster according to Scheme 1. We did not find any outlying rules that would satisfy Scheme 1 but did not have similar confidence.

The cluster representatives of other rules eliminated according to Scheme 1 are as follows:

CurrSch /exams.php ==> /ads.php , conf: 0.61

CurrSch /exams.php ==> /AnnualSch.php , conf: 0.36

CurrSch ==> /exams.php , conf: 0.83

CurrSch ==> /ads.php , conf: 0.60

CurrSch ==> /exams.php /ads.php , conf: 0.59

CurrSch ==> /AnnualSch.php , conf: 0.36

CurrSch ==> /AnnualSch.php /exams.php , conf: 0.34

All rules introduced as a cluster representative according to Scheme 1 are listed in the Appendix and can also be found at http://www.vtsns.edu.rs/maja/.

Pruning According to Scheme 2: Exchanging Sets of Rules by Their Cluster Representative

When pruning the rules in our sample data set, we identified 9 clusters of rules that fit into Scheme 2. We have eliminated 27 rules and introduced 9 rules as their cluster presentation (1 for each cluster), thus decreasing the size of the rule set by 18 rules.

For example, we have eliminated the following three rules and introduced their cluster representative. The confidence of all eliminated rules is close, which means they all fit into the Scheme 1.

/AnnualSch.php ==> /CurrSch.php /exams.php , conf: 0.69

/AnnualSch.php ==> /CurrSch.doc /exams.php , conf: 0.64

/AnnualSch.php ==> /CurrSch.doc /CurrSch.php /exams.php , conf: 0.64

The cluster representative of the three eliminated rules is:

/AnnualSch.php ==> /CurrSch /exams.php , conf: 0.66

Similar to the pruning according to Scheme 1, in our rule set all rules that followed the format of Scheme2 had a similar confidence. Therefore, we were able to eliminate three non-interesting rules for every rule-cluster according to Scheme2. We did not find any outlying rules that would satisfy Scheme2 but did not have similar confidence.

The cluster representatives of other rules eliminated according to Scheme2 are as follows:

/exams.php ==> CurrSch /ads.php , conf: 0.44

/ads.php ==> CurrSch /exams.php , conf: 0.28

/exams.php ==> CurrSch /AnnualSch.php , conf: 0.25

/AnnualSch.php /exams.php ==> CurrSch , conf: 0.74

/AnnualSch.php ==> CurrSch , conf: 0.69

/exams.php /ads.php ==> CurrSch , conf: 0.72

/exams.php ==> CurrSch , conf: 0.68

/ads.php ==> CurrSch , conf: 0.29

All rules introduced as a cluster representative according to Scheme 2 are listed in the Appendix and can also be found at http://www.vtsns.edu.rs/maja/.

Interestingness of the Resulting Association Rules

Pruning our rule set according to Schemes 1, 1a, and 2 decreased the size of the rule set from 96 to only 29 rules (more than 3 times).

We browsed through the rule set and identified those rules that we consider to be truly interesting to the user. We considered a rule truly interesting if, according to our knowledge of the website and its structure, we could identify an action that a webmaster can take in order to enhance the website structure and improve its browsing experience for the visitors. We identified 12 truly interesting rules out of the 29 rules in the rule set (41%).

In this section, we discuss those rules and why we consider them interesting. We divided the interesting rules into groups, which helps interpret their meaning and their potential usefulness to the user. Each group is listed in a separate table.

Table 1 shows two rules that contain two related but not directly linked items. Note that the item “CurrSch” represents two pages (CurrSch.php and CurrSch.doc), none of which is directly linked to the page AnnualSch.php.

Table 1: Rules related to annual and current exam schedules

No.	ITEM SET X	ITEM SET Y	CONFIDENCE
1.	AnnualSch.php	CurrSch	0.69
2.	CurrSch	AnnualSch.php	0.36

The first rule in Table 1 tells us that the visitors who are interested in viewing the annual exam schedule are also interested in viewing the current exam schedule in 69% of cases. The second rule tells us that the reverse is true in only 36% of cases. An example of an action that a webmaster might take based on this knowledge is to add a direct link from the page AnnualSch.php to the pages related to the current exam schedule, in order to make that information more easily accessible to the visitors.

Table 2 shows three rules that contain page exams.php on the left hand side and pages related to the current and annual schedule on the right hand side. The page exams.php contains direct links to the pages related to the current and annual schedule. As might be expected, the rules show that the visitors of exams.php are interested in the current exam schedule approximately 2 times more often than in the annual exam schedule. Furthermore, they are interested in both schedules 25% of the time. An example of an action a webmaster may take based on the rules in Table 2 is to emphasize the link to the current schedule on the page exams.php and list it before the link to the annual schedule (which was not the case at the time of this study).

Table 2: Rules with general exam page on the left hand side

No.	ITEM SET X	ITEM SET Y	CONFIDENCE
1.	/exams.php	CurrSch	0.68
2.	/exams.php	AnnualSch.php	0.33
3.	/exams.php	CurrSch /AnnualSch.php	0.25

Table 3 shows 3 rules, which all have the page ads.php on the right hand side and a similar value of confidence. It is important to note that none of the pages on the left hand side has a direct link to the page ads.php. The rules reveal the information that more than 60% of the visitors to the pages that contain information about exams (exam results, annual exam schedule, or current exam schedule) also visit the page ads.php. When looking at the exact values of confidence, we discovered that the visitors of the exam results page (examResults.php) were more interested in ads.php than those that visited the exam schedule (AnnualSch.php or pages in the CurrSch cluster).

Based on this knowledge, a webmaster may take various actions. For example, she/he may decide to add a link to ads.php on the examResult.php page. Alternatively, they may conduct further analyses to find out which particular ads that appear on the page ads.php would be worth adding directly the examResults.php page. Since the page ads.php contains links to various ads, this analysis might involve running an association rule or another rule-finding algorithm on a restricted set of pages that contain the page examResults.php and the pages of particular ads contained on ads.php. This may require finding association rules that have low support (lower than minimum), which is a specific problem addressed in the literature. We have left these further analyses out of the scope of this paper.

Table 3: Rules with ads page on the right hand side

No.	ITEM SET X	ITEM SET Y	CONFIDENCE
6.	examResults.php	ads.php	0.66
7.	AnnualSch.php	ads.php	0.62
8.	CurrSch	ads.php	0.60

Table 4 shows 3 rules which all have the page ads.php on the left hand side. Page ads.php does not have a direct link to any of the pages on the right hand side. According to these rules, between 30% and 40% of the visitors of the page ads.php also visit the pages about exams (29 % current schedule, 27% exam results, 39% general exam information). A webmaster could consider the link structure between those pages and might decide to either make some changes or leave it as is, since the percentages are not as high and adding additional links may be a burden on the page ads.php, especially if it is already crowded.

Table 4: Rules with ads page on the left hand side

No.	ITEM SET X	ITEM SET Y	CONFIDENCE
9.	ads.php	exams.php	0.39
10.	ads.php	CurrSch	0.29
11.	ads.php	examResults.php	0.27

Table 5 shows the only rule in our rule set that has page ads.php on the left hand side and a page to which there is a direct outgoing link on the page ads.php. It is important to note that there are many links on the page ads.php, but those pages either did not have high enough support or the rules did not have high enough confidence.

Table 5: Rules with direct outgoing links on the ads page

No.	ITEM SET X	ITEM SET Y	CONFIDENCE
12.	ads.php	ExamRegNov.doc	0.17

Based on this rule, a webmaster may decide to emphasize the link to ExamRegNov.doc on the page ads.php. Or, since it seems that ExamRegNov.doc is rather important document on ads.php compared to other links on that page, a webmaster might decide to do further analysis (possibly running a version of association rule finding with lower minimum support and targeting ExamRegNov.doc attribute), in order to find out whether the page ExamRegNov.doc is correlated with some other pages. If such correlations are found, the webmaster might add links to ExamRegNov.doc on those pages.

Conclusion

We conducted a comprehensive analysis of web usage association rules found in a website of an educational institution. Our experiments confirmed that one of the major issues in association rule finding is the existence of too many rules, all of which satisfy defined constraints, but it is difficult to exploit and identify those that are truly interesting to the user.

Our experiments confirmed that, particularly in the domain of web usage mining, the size of association rule set increases dramatically due to the existence of “hard” rules that have very high confidence due to the interconnectedness of web pages through the link structure.

In order to deal with the issue of rule over-generation, we proposed and implemented a set of schemes for pruning irrelevant rules, which are merely a consequence of the “hard” rules. The schemes are simple and straightforward and can serve as a basis for developing a software tool for pruning association rule set generated by Weka (Weka, n.d3), or any other data mining software.

Using the knowledge of the website structure and the behavior of the site’s visitors, we analyzed the pruned rule set from the user’s point of view and proposed actions that a webmaster may decide to take based on knowledge extracted from the rules in order to enhance a website and improve visitors' browsing experience.

The percentage of truly interesting rules that can prompt a webmaster to actions that improve the structure of a website compared to the whole rule set size in our experiment was 41%. This confirms our hypothesis that browsing for the truly interesting rules in the rule set can be a useful task worth undertaking.

Directions for Future Work

Rule pruning schemes for "transitory" rules

Similar to the rule pruning schemes related to page clusters (Definition 1), it is also possible to devise pruning schemes based on the rules of the form a==>b, conf(a==>b) ≈ 1 , when the rule b==>a does not hold or doesn't have confidence close to 1. We call the rules that could be pruned out based on such schemes "transitory".

In our experiment, rule pruning based on page clusters as described in the previous section performed the major part of pruning the rule set, making the resulting rule set easy to browse for truly interesting rules. While this additional pruning may further simplify browsing the rule set, it was not necessary for our case study and we left it out of the scope of this paper.

Additional pruning techniques

Within the domain of mining association rules in the web usage log data there is a need to identify specific pruning techniques as well as interestingness measures that, in particular, address the issue of web page interconnectedness through the website link structure. Such measures could, for example, be based on the statistical expectedness of the confidence of an association rule according to the link structure of the website, when compared to the actual confidence of the rule according to data mining results. We will leave this research direction for future work.

Larger data sets

The pruning schemes that we used proved to be efficient in decreasing the rule set in our example to a manageable size by producing a very high percentage of truly interesting rules. However, we cannot claim that for larger websites and larger data sets that will also be the case. The initial pruning can still be based on the same schemes, while other pruning methods may be applied to further decrease the rule set. We plan to undertake further research and investigate how far it is necessary to go until the rule set size has a manageable number of interesting rules for larger websites and data sizes.

Software implementation

According to our experimental results, a software application that applies the pruning techniques presented in this paper and possibly extended by additional pruning techniques would be a valuable ad-on to the Weka (Weka, n.d.) data mining software. To our knowledge, this has not yet been implemented nor is it under development at Weka. Additionally, while researching the set of generated rules in our experiment and looking for those that are truly interesting, we found that another valuable addition to Weka, or other association rule mining tools, would be a tool that helps organize the set of all rules (perhaps in related clusters, trees, or some other structures) and help browse through them. We suggest this for future studies.

References

Aggarwal, C. C., & Yu, P. S. (1998). A new framework for itemset generation. In *PODS 98, Symposium on Principles of Database Systems*, pages 18-24.

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

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the International Conference on Very Large Databases,* Santiago, Chile, pp. 478-499. Los Altos, CA: Morgan Kaufmann.

Anand, S. S., Mulvenna, M., & Chavielier, K. (2004). On the deployment of web usage mining. In B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, & G. Stumme (Eds.), *Web mining: from web to semantic web* (pp. 23-42). Berlin: Springer.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255-264, 265-276.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the world wide web. Proceedings of the *IEEE International Conference. Tools with AI,* Newport Beach, CA, pp. 558-567.

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.

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

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.

Iváncsy, R., & Vajk, I. (2008). Frequent pattern mining in web log data. *Journal of Applied Sciences at Budapest Tech, 3*(1), Special Issue on Computational Intelligence.

Jaroszewicz , S., & Simovici, D. A. (2002). Pruning redundant association rules using maximum entropy principle. *Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference, PAKDD'02*.

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

Liu, B., Hsu, W., & Ma, Y. (1999). Pruning and summarizing the discovered associations. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 125-134.

Meo, R., Luca Lanzi, P., Matera, M., Careggi, D., & Esposito, R. (2004). Employing inductive databases in concrete applications, constraint-based mining and inductive databases. *European Workshop on Inductive Databases and Constraint Based Mining.*

Omiecinski, E.R. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering, 15*(1):57-69.

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

Wang Y., Li, Z., & Yang, Z. (2005). Mining sequential association-rule for improving Web document prediction. *Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications,* pp. 146 – 151.

Weka. (n.d.) Data mining software in Java. The University of Waikato. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/

Appendix

Association rules after the pruning according to the Schemes 1, 1a and 2.

/ExamRegNov.doc ==>	/ads.php	conf:	1
/CurrSch.doc ==>	/CurrSch.php	conf:	0.99
CurrSch /AnnualSch.php ==>	/exams.php	conf:	0.93
/AnnualSch.php ==>	/exams.php	conf:	0.9
CurrSch ==>	/exams.php	conf:	0.85
CurrSch /ads.php ==>	/exams.php	conf:	0.91
/CurrSch.php ==>	/CurrSch.doc	conf:	0.85
/AnnualSch.php /exams.php ==>	CurrSch	conf:	0.74
/AnnualSch.php ==>	CurrSch	conf:	0.69
/exams.php ==>	CurrSch	conf:	0.68
/examResults.php ==>	/ads.php	conf:	0.66
/AnnualSch.php ==>	CurrSch /exams.php	conf:	0.66
/AnnualSch.php ==>	/ads.php	conf:	0.62
CurrSch /exams.php ==>	/ads.php	conf:	0.61
/exams.php ==>	/ads.php	conf:	0.61
CurrSch ==>	/ads.php	conf:	0.6
CurrSch ==>	/exams.php /ads.php	conf:	0.59
/exams.php ==>	CurrSch /ads.php	conf:	0.44
/ads.php ==>	/exams.php	conf:	0.39
/examResults.php ==>	/exams.php	conf:	0.38
CurrSch /exams.php ==>	/AnnualSch.php	conf:	0.36
CurrSch ==>	/AnnualSch.php	conf:	0.36
CurrSch ==>	/AnnualSch.php /exams.php	conf:	0.34
/exams.php ==>	/AnnualSch.php	conf:	0.33
/ads.php ==>	CurrSch	conf:	0.29
/ads.php ==>	CurrSch /exams.php	conf:	0.26
/ads.php ==>	/examResults.php	conf:	0.27
/exams.php ==>	CurrSch /AnnualSch.php	conf:	0.25
/ads.php ==>	/ExamRegNov.doc	conf:	0.17

Biographies

Maja Dimitrijević is a lecturer at the Advanced Technical School in Novi Sad. She teaches 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, and has 5 years of experience in software development.

Zita Bošnjak 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.