DATA ADVANCE PREPARATION FACTORS AFFECTING RESULTS OF SEQUENCE RULE ANALYSIS IN WEB LOG MINING

Michal Munk, Jozef Kapusta, Peter Švec, Milan Turčáni

Introduction

Currently, Internet represents a most dynamically developing source of information, thus becoming an important data source. Web pages are first of all understood as a source of information for users - clients. Only a few companies are conscious of the fact that web pages can provide information also in the opposite direction. Organization can gain a number of information on its clients, on their behaviour, on their interests, etc.. As a result of the need to analyze these data a related area of the knowledge discovery in databases (hereinafter referred to as KDD), the so-called web mining, originated.

One of the methods of web mining is also discovering patterns of behaviour of portal visitors. Based on the found users behavior patterns, which are represented by sequence rules, it is possible to modify and improve web page of the organization. The data preparation itself represents the most time consuming phase of the web page analysis. Data for the analysis are gained from the web server log file.

The aim of the article is to find out using an experiment to what measure it is necessary to execute data preparation for web log mining and determine inevitable steps for gaining valid data from the log file. Importance of opportunities of data advance preparation through a log file analysis was verified by means of experiment. We were interested in finding out which steps are important in terms of a correct analysis of portal for anonymous visitors. Results of the experiment are very important for the portal, which is regularly analysed and modified, since they prove correctness of individual steps upon analysing, or they can make the data advance preparation more simple through an identification of "useless" steps.

KDD is aimed in finding new, interesting and useful knowledge through utilising statistical methods and computer learning methods for the purpose of classification [19, 36], segmentation [15], association rules discovering [1], sequential rules discovering [11], etc.. In case that these data are mined from web, we call this process web mining. Web mining can be defined as an extraction of interesting and potentially useful knowledge and information from activities referring to World Wide Web [16].

Sometimes, for the application of web mining methods, it is sufficient only to slightly adjust the existing procedures from the scope of KDD, otherwise it is necessary to change steps of data advance preparation and transformation more radically.

Web mining can be divided into three domains [33]:

a. knowledge discovering based on the web contents (web content mining),
b. knowledge discovering based on the web structure (web structure mining),
c. knowledge discovering based on the web usage (web usage mining).

Web usage mining (hereinafter referred to WUM) is focused on the analysis of behaviour of users while surfing the net [4, 25, 27]. The most frequent sources of data are the ones automatically stored in the log files. This is the reason why this area is often marked as web log mining. In such data we follow series – sequences in visiting individual pages by the user, who is, under certain condition, identified by the IP address. In sequences we can look for users behaviour patterns. For this purpose it is the best way to use sequence rule analysis, the aim of which is to extract sequence rules. By means of these rules sequences of visits of various web sections by the user are predicted. This method was deduced from association rules and can serve as an example of the method making provision for

Internet peculiarities.

Association rules and cluster analysis can serve as examples of the use of standard methods of data mining upon web access analysing. For example, if multiple stores issue customer cards in order to obtain information on their customers (by means of which they can get discounts), they get records on customers along with the shopping linkage, then, for electronic shops, which obtain information upon customer registration, it should be even more simple. Through segmentation, for instance, we can examine behaviour of groups of clients using cluster analysis and consequently we can define the obtained segments by association rules.

The most frequent applications of web log mining are as follows [20]:

a. usage analysis,
b. web optimization,
c. web personalization.

Good quality data are a prerequisite for a well--realized data analysis. If there is "junk" at the input, the same will be at the output, regardless of the method for knowledge extraction used. This applies even more in the area of web log mining, where the log file requires a thorough data preparation. As an example we can present the usage analysis, where we are aimed at finding out what our web visitors are interested in.

For this purpose we can use:

a. survey sampling – we find out answers to particular items in the questionnaire and a visitor of our site knows that he/she is the object of our survey [10],
b. web log mining – we analyse the log file of the web server, which contains information on accesses to the pages of our web, and the visitor does not know that he is the object of our survey [8].

While in case of the survey sampling we can provide good quality data using a reliable and valid measuring procedure for their mining, in case of the web log mining we can provide them through good preparation of data from the log file. In its standard structure called Common Log File [32] it records each transaction, which was executed by the browser at each web access. Each line represents a record with the IP address, time and date of the visit, accessed object and referenced object. In case that we will use its extended version, we can record user browser version, the so-called User-Agent.

Log file of the web server is a source of anonymous data about the user. These anonymous data represent also the problem of unique identification of the web visitor. Reconstruction of activities of each visitor is demanding. Currently it is common that several users share a common IP address, whether they are situated under a certain NAT (Network Address Translation), or proxy equipment. Authentication mechanisms can facilitate identification of the user, however, their usage is undesirable due to privacy protection [3]. Another problem which WUM should face, are crawlers of various search engines, which browse through the whole web, mostly recursively and successively. Detection of crawlers is possible either based on their identification by means of the User-Agent field, or IP address by their comparison with the www.robotstxt.org database. This database need not contain data on all crawlers, however, those minority ones represent a neglectable number. Another method of identification is to find, whether crawler accessed to the file robots.txt or not [17]. Based on the access to this file we can unambiguously identify the crawler even if his User-Agent array is incorrectly set.

One of the possibilities how to differentiate individual visitors is to do it on the various versions of the Internet browser [8]. We can expect that if there exist several accesses from a single IP address with various versions of the browser or operating system, there is not only one user [24]. Cooley et. al. [8] also assume that if an access to the page, which is not accessible from the previous page, has been recorded, such an access can be accessed as the one by other user. This observation, however, is not unequivocal, since the user can run records on his/her favourite items and thus access also such subpages, which are not referenced from the previously accessed page [3].

Individual visitors can be differentiated also based on the identification of sessions. The aim of sessions identification is to divide individual accesses of each user into separate relations [8]. These relations can be defined in various ways. Session can be defined as a series of steps, which lead to the fulfilment of a certain task [26], or as a series of steps, which lead to the reaching of a certain goal [7]. The simplest method is to consider a session to be a series of clicks for a certain period of time, e.g. 30 minutes [3]. A real value for session can be derived based on empirical data.

Another problem of WUM is a reconstruction of activities of a web visitor. Taucher and Greenberg [30] proved that more than 50 % of accesses on web are backward. This is the beginning of the problem with the browser´s cache. In backward, no query to web server is executed, so there does not exist any record of it in the log file. One of the solutions of this problem is path completion, through which we add these missing records to the log file [8].

A remarkable application of WUM is its use for system performance analysis [14], optimum cache creation [5, 9], identification of the most suitable place for an advertisement [6], building of adaptive web pages [23, 25]. Generally applications using WUM techniques can be divided into five categories: personalization, systems improvement, web pages modification, bussines intelligence and characteristics of utilisation [27]. Park et al. [22] present numerous studies, where technologies such as booleous vector models, frequency representations of web usage, feature-matrices (FM) for access patterns mining, self-organizing maps, were used for identification of users sessions. All these studies group users based on similar characteristics. Another approach is looking for sequential rules [12, 13, 18, 21].

All these techniques depend on the primal step of gathering and cleaning of data, as it is described by Cooley et. al. [8]. In our experiment we tried to find out what steps were necessary when using sequence rule analysis.

1. Techniques and Methods Used

1.1 Data Preparation Techniques

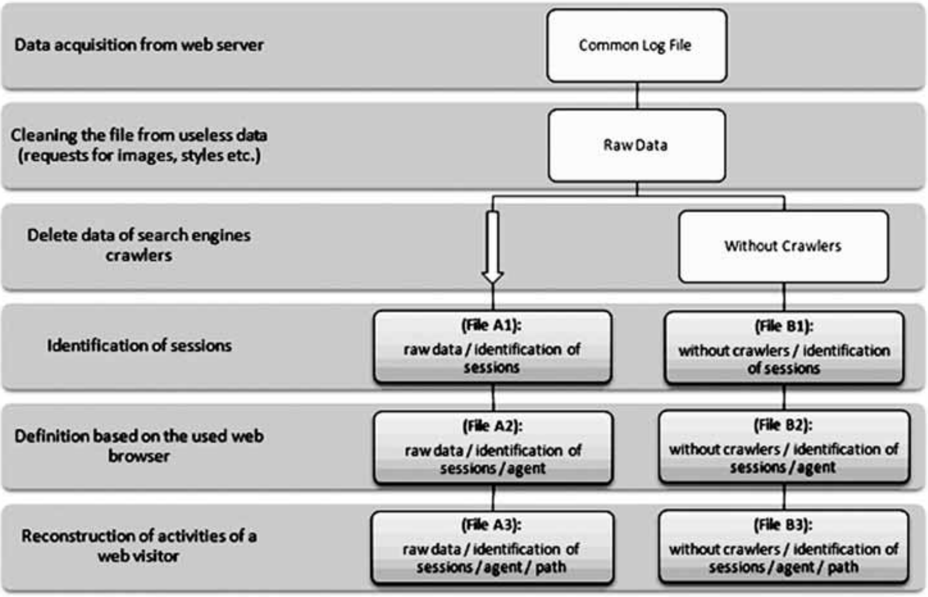
The data preparation for the needs of our experiment consisted of the following steps.

1. The first step in the log file adjustment was **cleaning the file** from useless data. Under useless data we understand mainly the lines of the log file, in which are recorded requests for images, styles and scripts or other files, which can be inserted into the page. This part is the most simple from the whole data preparation process, since it consists of only filtration of the data, which do not comply with the selected template. A file of raw data of the log file resulted from this step.
2. The pages are accessed also by **crawlers** of various **searching engines,** which proceed in a different way than a common visitor does. By means of a simple detection of these crawlers and their deletion from the log file the resulting file with accesses only from standard visitors (without crawlers) was obtained.
3. Another step in the advance preparation of data was **identification of sessions**. Identification of sessions of the user allows us to eliminate NAT and proxy devices, as well as identify various users alternating behind one computer. From our point of view sessions were identified as a delimited series of clicks realized in the defined period of time. In spite of the recommended 30-minute-long time window we chose the 10 minute time window with regard to the variable avg. time on site obtained by means of the Google Analytics tool, which represents average time of the user on our web page. While the current steps were the concern of simple programmes, which sequentially scanned Common Log File, upon identificating the sessions it was necessary to enter all data from the file into the database, with which our application for the advance preparation of data for sequence rule analysis cooperated. After applicating the algorithm ensuring identification of sessions to the file of raw data and the file cleaned from crawlers we obtained files (File A1) - raw data/identification of sessions and (File B1) - without crawlers/identification of sessions.
4. One of the methods of identification of users (hiding behind various NAT devices or proxy servers) is their definition based on the **used web browser,** i.e. records from identical IP address were more specifically divided into individual sessions as to the used browser. This way we can specify also the sessions of users from Internet cafés, computer classrooms, etc., where several users alternate behind one computer, and we assume that not all of them use the same web browser. The result of this modification was the files (agents) with a closer division of users sessions, i.e. (File A2) - raw data/identification of sessions/agent and (File B2) – without crawlers/identification of sessions/agent.

5. Another problem upon searching for the users behaviour patterns seems to be the analysis of the backward path, or **reconstruction of activities of a web visitor.** Reconstruction of activities is focused on retrograde completion of records on the path went through by the user by means of a **back button**, since the use of such button is not automatically recorded into the Common Log File. A sitemap has a great importance for retrograde completion of the path. We can find in it information on the existence of a link among pages, i.e. if a hyperlink from one page to another exists. The sitemap was obtained for the needs of our analysis by means of Web Crawling application implemented in the used Data Miner. Having lined up the records according to the IP address we searched for some linkages between the consecutive pages. A sequence for the selected IP address can look like this: A→B→C→D→X. Based on the sitemap the algorithm in our example can find out that there does not exist the hyperlink from the page D to our page X. We thus assume that this page was accessed by the user by means of using a Backbutton from one of the previous pages. Through a backward browsing we then find out, on which of the previous pages exists a reference to page X. In our sample case we can find out that if there does not exist a hyperlink to page X from page C, if C page is entered into the sequence, i.e. the sequence will look like this: A→B→C→D→C→X. Similarly, we shall find that there does not exist any hyperlink from page B to page X and add it into the sequence, i.e. A→B→C→D→C→B→X. Finally algorithm finds out that page A contains hyperlink to page X and after the termination of the backward path analysis the sequence will look like this: A→B→C→D→C→B→A-→X. It means then, that the user used Back button in order to transfer from page D to C, from C to B and from B to A. After the application of this method to files (File A2) and (File B2) we obtained files (File A3) - raw data/identification of sessions/agent/path and (File B3) - without crawlers/identification of sessions/agent/path.

For the needs of our experiment all these steps were taken using independent applications, i.e. we can simply say that a separate application was created for each step.

Fig. 1: Application of data advance preparation to the log file


Data acquisition from web server
Common Log File
Cleaning the file from useless data (requests for images, styles etc.)
Raw Data
Delete data of search engines crawlers
Without Crawlers
Identification of sessions
(File A1): raw data / identification of sessions
(File B1): without crawlers / identification of sessions
Definition based on the used web browser
(File A2): raw data / identification of sessions / agent
(File B2): without crawlers / identification of sessions / agent
Reconstruction of activities of a web visitor
(File A3): raw data / identification of sessions / agent / path
(File B3): without crawlers / identification of sessions / agent / path

Source: our own research

The following scheme (Fig. 1) shows an overview of application of individual methods on Common Log File and the following creation of files for the analysis.

1.2 Sequence Rules

Sequence rules have been derived from association ones, thus the differences are not so wide [31]. k – sequence is the one of the length k, i. e. it contains k pages. Frequented k – sequence is a variation of the frequented k – item set, or a combination. Usually, the frequented single-item set is identical with the frequented single sequence.

In the following example we shall illustrate differences between the algorithm Apriori and AprioriAll, where the algorithm Apriori serves for the searching of association rules and AprioriAll for the searching of sequence rules: *D* = {S1 = {*U*1, <*a, b, c*>}, S2 = {*U*2, <*a, c*>}, S3 = {*U*1, <*b, c, e*>}, S4 = {*U*3, <*a, c, d, c, e*>}}, where D is a databasis of transactions with a time label, a, b, c, d, e are web sections and U1, U2, U3 represent users. Each transaction is identified by a user. Let us assume that the minimum support is 30 %.

In this case, user U1 has actually two transactions. When searching for sequence rules we consider his sequence to be a current connection of web sections in the transactions S1 and S3, i.e. a sequence can consist of several transactions, while continuous accesses to pages are not required. Similarly support of the sequence is designated not by the percentage of transactions, but by the percentage of users, who own the given sequence. Sequence is large/frequented, if it is at least situated in one sequence identified by the user and meets the condition of minimum support. Set of frequented sequencies, which have k items, we mark *Lk*. For finding *Lk*, we use set of candidates, marks *Ck*, which involves sequences with k items.

The first step is ranking of transactions as to the user with a time label of each page visited by him; the remaining steps are similar to the ones in algorithm Apriori. Having ranked the transactions we obtained current sequences identified by the user, which represent complete references from a single user: D = {*S*1 = {*U*1, <*a, b, c*>}, S3 = {*U*1, <*b, c, e*>}, S2 = {*U*2, <*a, c*>}, S4 = {*U*3, <*a, c, d, c, e*>}}. Similarly to algorithm Apriori we start with generating the set of candidates of length 1: *C*1 = {<*a*>, <*b*>, <*c*>, <*d*>, <*e*>}, from it we define the set L1 = {<*a*>, <*b*>, <*c*>, <*d*>, <*e*>} of single-item sequences, where each page is referenced at least by one user. Followingly, we generate sets of candidates *C2* from *L1* by means of the so-called full linking, i.e. we make provisions for the web user who searches through the pages forwards or backwards. This is the reason why algorithm Apriori is not suitable for web log mining, by contrast to algorithm AprioriAll, which respects the above mentioned fact. Out of the set of candidates of length 2: *C2* = {<*a, b*>, <*a, c*>, <*a, d*>, <*a, e*>, <*b, a*>, <*b, c*>, <*b, d*>, <*b, e*>, <*c, a*>, <*c, b*>, <*c, d*>, <*c, e*>, <*d, a*>, <*d, b*>, <*d, c*>, <*d, e*>, <*e, a*>, <*e, b*>, <*e, c*>, <*e, d*>}, we shall define a set *L2* = {<*a, b*>, <*a, c*>, <*a, d*>, <*a, e*>, <*b, c*>, <*b, e*>, <*c, b*>, <*c, d*>, <*c, e*>, <*d, c*>, <*d, e*>} of double-item sequences, where each sequence is situated in at least one sequence identified by the user. We then analogically proceed.

2. Experiment

Experiment was realized in several steps.

1. Data acquisition – defining the observed variables into the log file from the point of view of obtaining the necessary data (IP address, date and time of access, URL address, etc.).
2. Creation of data matrices – from the log file (information of accesses) and sitemaps (information on the web contents).
3. Data preparation (identification, transformation, cleaning of data, etc.) on various levels:
   AY – raw data cleaned only from unnecessary data,
   BY – cleaned data from the accesses of search engines crawlers,
   X1 – with an identification of sessions,
   X2 – with an identification of sessions an agent allowance,
   X3 – with an identification of sessions with making provisions for the agent and completing the paths, where X = {A,B} a Y = {1,2,3}).

   Combining these levels of data preparation we obtain the following files:
   a. (File A1): raw data/identification of sessions,
   b. (File A2): raw data/identification of sessions/agent,
   c. (File A3): raw data/identification of sessions/agent/path,

   d. (File B1): without crawlers/identification of sessions,
   e. (File B2): without crawlers/identification of sessions/agent,
   f. (File B3): without crawlers/identification of sessions/agent/path.
4. Data analysis – searching for behaviour patterns of web users in individual files.
5. Understanding the output data – creation of data matrices from the outcomes of the analysis, defining assumptions.
6. Comparison of results of data analysis elaborate on various levels of data preparation from the point of view of quantity and quality of the found rules – patterns of behaviours of users upon browsing the web:
   a. comparison of the portion of the rules found in examined files,
   b. comparison of the portion of inexplicable rules in examined files,
   c. comparison of values of the degree of support and confidence of the found rules in examined files.

Uncleaned file (Tab. 1) containes almost 40000 cases, which represent accesses to portal during one week, of which almost 11 % of cases are crawlers accesses. Having cleaned the files from crawlers the number of visits (costumer's sequences) decreased by approximately 10 % and on the contrary, the number of frequented sequences increased by 11 % to 19 % in examined files. Having completed the paths the number of records increased by almost 70 % and the average length of sequences increased from 4 to 7, or to 6 in case of files cleaned from crawlers accesses. On the other hand, upon making provision for the used browser (agent) when identifying sessions we can follow only about 4 % growth of visits (costumer's sequences).

Based on these primary results we articulated the following assumptions:

a. we assume that cleaning of data from crawlers accesses will not have a significant impact on the quantity of extracted rules but, on the contrary, on their quality in terms of reducing the portion of inexplicable rules,
b. we expect that completion of paths will have a significant impact both on the quality and quantity of extracted rules,
c. we assume that making provisions for the used browser upon identifying sessions will not have a significant impact on the results of the analysis.

2.1 Comparison of the Portion of the Found Rules in Examined Files

Users´ accesses to individual web sections of the university portal were observed in the course of one week. The analysis (Tab. 2) resulted in sequence rules, which we obtained from frequented sequences fulfiling their minimum support (in our case min s = 0.03). Frequented sequences were obtained from identified sequences, i.e. visits of individual portal users during one week.

There is a high coincidence between the results (Tab. 2) of sequence rule analysis in terms of the portion of the found rules in case of files without completing paths (X1, X2) in both groups (AY, BY). The most rules were extracted from

Tab. 1: Number of accesses and sequences in particular files

	Count of web accesses	Count of costumer's sequences	Count of frequented sequences	Average size of costumer's sequences
File A1	39688	9904	58	4
File A2	39688	10285	57	4
File A3	67236	10285	69	7
File B1	35374	8875	69	4
File B2	35374	9259	63	4
File B3	60087	9259	80	6

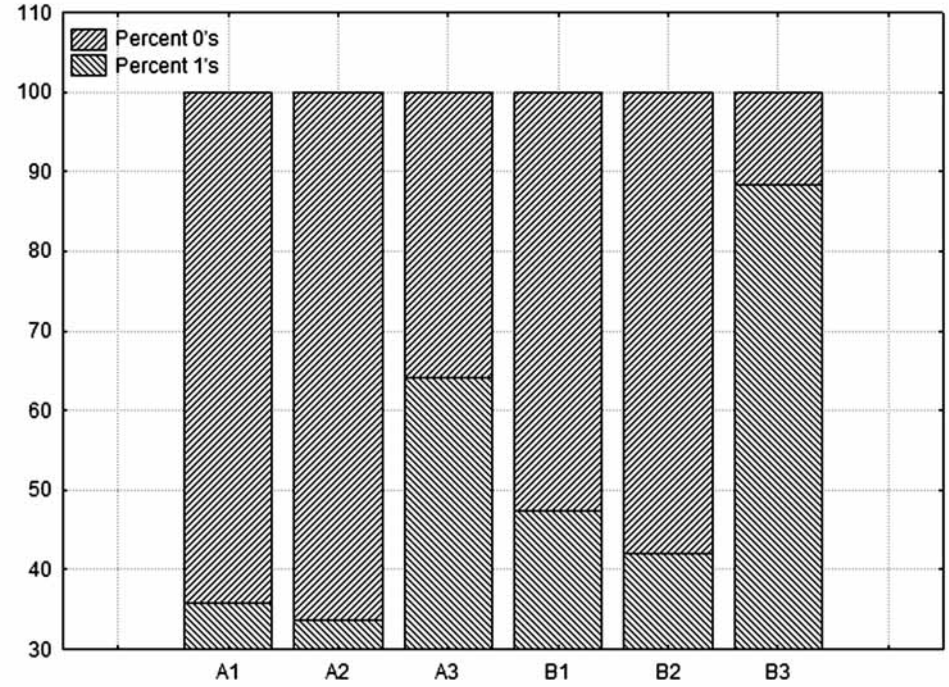
Source: our own research

Tab.2: Discovered sequence rules in particular files

Body	==>	Head	A1	A2	A3	B1	B2	B3
(a178)	==>	(a180)	1	1	1	1	1	1
⋮	==>	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(c3), (a180)	==>	(a178)	1	0	0	1	1	0
(c3), (a49)	==>	(a369)	1	1	0	1	1	1
⋮	==>	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(c7)	==>	(a64)	1	1	1	1	1	1
Count of derived sequence rules			34	32	61	45	40	84
Percent of derived sequence rules (Percent 1's)			35.8	33.7	64.2	47.4	42.1	88.4
Percent 0's			64.2	66.3	35.8	52.6	57.9	11.6
Cochran Q test			Q = 118.2868, df = 5, p < 0.001					

Source: our own research

Fig. 2: Sequential/Stacked plots for derived rules in examined files



Source: our own research

Tab. 3: Homogeneous groups of examined files

File	Mean	1	2	3
A2	0.337	****		
A1	0.358	****		
B2	0.421	****		
B1	0.474	****		
A3	0.642		****	
B3	0.884			****

Source: our own research

files with completing paths, concretely 61 were extracted from the file A3, which represents over 64 % and 84 from the file B3, which represents over 88 % of the total number of found rules. Generally, more rules were found in the observed files cleaned from the crawlers accesses (BY).

The following graph visualizes (Fig. 2) the results of Cochran´s Q test.

Based on the results of Q test (Tab. 2), the zero hypothesis, which reasons that the incidence of rules does not depend on individual levels of data preparation for web log mining, is rejected at the 1 % significance level.

From the multiple comparison (Tukey Unequal N HSD test) (Tab. 3) a single homogenous group consisting of files A1, B1, A2 and B2 was identified subject to the average incidence of the found rules. Statistically significant differences on the level of significance 0.05 in the average incidence of found rules were proved among file A3 and the remaining ones, B3 and the remaining ones, and between files A3 and B3.

If we have a look at the results in details (Tab. 4), we can see that in the files cleaned from crawlers (BY) were found identical rules to the files containing raw data (AY), while the difference consisted only in 8 to 23 new rules, which were found in the files cleaned from crawlers (BY). In case of files without completion of paths (B1, B2) the portion of new files represented 8 % and 12 %. In case of the file with the completion of the path (B3) more than 24 %, where also the statistically significant difference in the number of found rules between A3 and B3 in favour of B3 was proved.

Cleaning of data from the crawlers accesses has an important impact on the quantity of extracted rules only in case of files with the completion of paths (A3 vs. B3). It also has been proved that completion of paths has an important impact on the quantity of extracted rules (X3 vs. X1, X2, X3). On the contrary, making provisions for the used browser upon identifying sessions has no significant impact on the quantity of extracted rules (A1, A2, B1, B2).

Tab. 4: Crosstabulations: 2 by 2 Tables – AY x BY

A1\B1	0	1	Total	A2\B2	0	1	Total	A3\B3	0	1	Total
0	50	11	61	0	55	8	63	0	11	23	34
	52.6 %	11.6 %	64.2 %		57.9 %	8.4 %	66.3 %		11.6 %	24.2 %	35.8 %
1	0	34	34	1	0	32	32	1	0	61	61
	0.0 %	35.8 %	35.8 %		0.0 %	33.7 %	33.7 %		0.0 %	64.2 %	64.2 %
Total	50	45	95	Total	55	40	95	Total	11	84	95
	52.6 %	47.4 %	100 %		57.9 %	42.1 %	100 %		11.6 %	88.4 %	100 %

Source: our own research

2.2 Comparison of the Portion of Inexplicable Rules in Examined Files

We require from association rules that they be not only clear but also useful. Association analysis produces three elementary kinds of rules [29]:

a. utilizable (useful, beneficial),
b. trivial,
c. inexplicable.

In our case upon sequence rules it is useless to consider trivial rules. We will differentiate only the utilizable and inexplicable rules.

The portion of inexplicable rules (Tab. 5) among the files cleaned from crawlers (BY) and the ones containing raw data (AY) is approximately identical on all three levels of data preparation (X1, X2, X3).

Phi-square represents the degree of relationship between two dichotomic variables (Rule, File). The value of coefficient (Tab. 5) is 0 in all three cases, while 1 means perfect relationship and 0 no relationship.

Cleaning of data from crawlers accesses has no impact on the reduction of portion of inexplicable rules.

2.3 Comparison of the Values of Support and Confidence Rates of the Found Rules in Examined Files

Quality of sequence rules is assessed by means of two indicators [29]:

a. support,
b. confidence.

Results of the sequence rule analysis showed differences not only in the quantity of the found rules, but also in the quality in terms of the support characteristics values of the discovered rules among individual files. Kendall´s coefficient of concordance represents the degree of concordance in the support of the found rules among examined files. In both groups (Fig. 3) the value of coefficient is approximately 0.35, while 1 means a perfect concordance and 0 represents discordancy. Results are visualized in line plots of multiple variables. Graphs visualize values of support of the found rules in individual files. Support values (Fig. 3) are not copied, which only proves the found discordancy in the support values of the found rules among individual files.

Correlation matrix (Tab. 6) shows that the largest degree of concordance/dependance in the support is between the rules found in the files with identification of sessions without allowance (X1) and with allowance of the agent (X2), regardless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY). On the contrary, a considerably smaller concordance is among files with completing paths (X3) and the remaining files, also regardless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY).

Results are visualized (Fig. 4) in the matrix plot, where an almost perfect concordance in support values of the found rules among files (A1, A2, B1, B2) and a markedly smaller, but still significant ($p < 0.05$), concordance among files with completing paths (A3, B3) and the others was found.

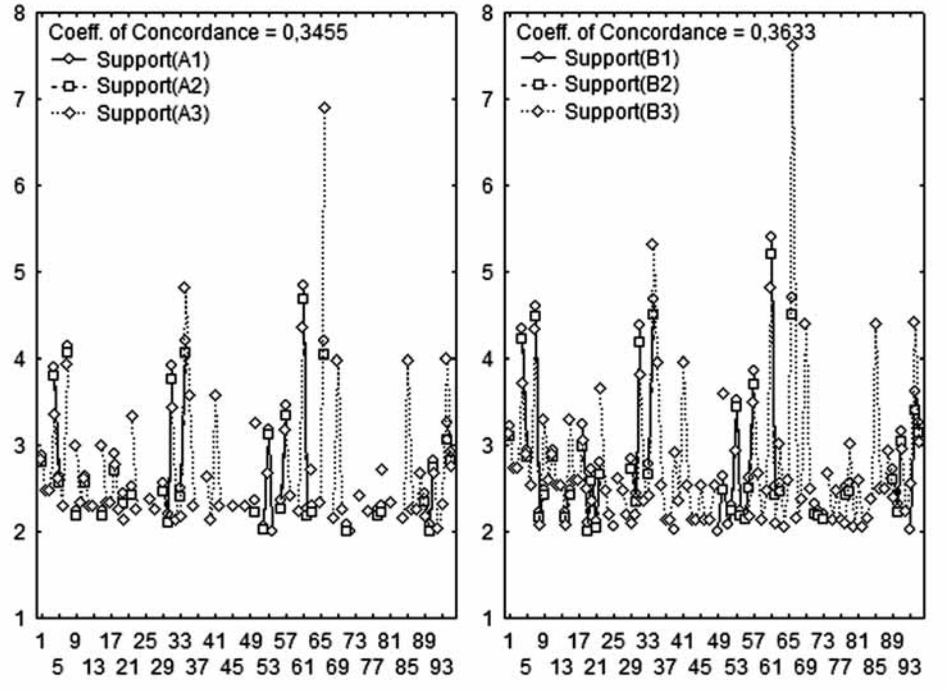
There were also demonstrated differences in

Tab. 5: Crosstabulations: 2 by 2 Tables – Rule x File

Rule\File	A1	B1	Rule\File	A2	B2	Rule\File	A3	B3
Utilizable	33	43	**Utilizable**	31	38	**Utilizable**	41	59
	97.1 %	95.6 %		96.9 %	95.0 %		67.2 %	70.2 %
Inexplicable	1	2	**Inexplicable**	1	2	**Inexplicable**	20	25
	2.9 %	4.4 %		3.1 %	5.0 %		32.8 %	29.8 %
Total	34	45	**Total**	32	40	**Total**	61	84
	100 %	100 %		100 %	100 %		100 %	100 %
Phi-square	0.00152		**Phi-square**	0.00217		**Phi-square**	0.00104	

Source: our own research

Fig. 3: Line plots of multiple variables for support of derived rules



Source: our own research

the quality in terms of confidence characteristics values of the discovered rules among individual files. In both groups (Fig. 5) the coefficient of concordance values is almost 0.2, while 1 means a perfect concordance and 0 represents discordancy. Results are visualized in line plots of multiple variables. Graphs visualize values of confidence the found rules in individual files. Confidence values (Fig. 5) are not copied, which only proves the found discordancy in the confidence values of the found rules among individual files.

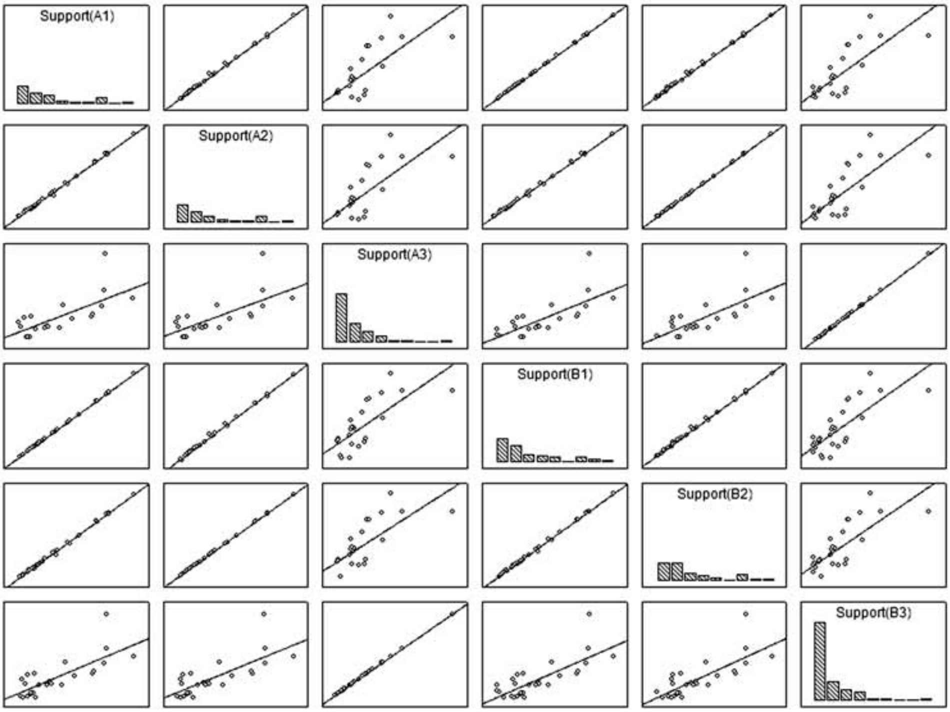
Correlation matrix (Tab. 7) shows that the largest degree of concordance/dependence in the confidence is among the rules found in the file with identification of sessions without allowance (X1) and with allowance of the agent (X2), regar-

Tab. 6: Kendall Tau correlations for support of derived rules

	Sup (A1)	Sup (A2)	Sup (A3)	Sup (B1)	Sup (B2)	Sup (B3)
Sup (A1)	1.000000	0.957073	0.487179	0.999100	0.963905	0.538655
Sup (A2)	0.957073	1.000000	0.455047	0.958047	0.993842	0.524470
Sup (A3)	0.487179	0.455047	1.000000	0.517241	0.488215	0.997200
Sup (B1)	0.999100	0.958047	0.517241	1.000000	0.975991	0.540541
Sup (B2)	0.963905	0.993842	0.488215	0.975991	1.000000	0.575821
Sup (B3)	0.538655	0.524470	0.997200	0.540541	0.575821	1.000000

Source: our own research

Fig. 4: Matrix plot for support of derived rules



Source: our own research

dless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY). On the contrary, a considerably smaller concordance is among files with completing paths (X3) and the remaining files, also regardless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY).

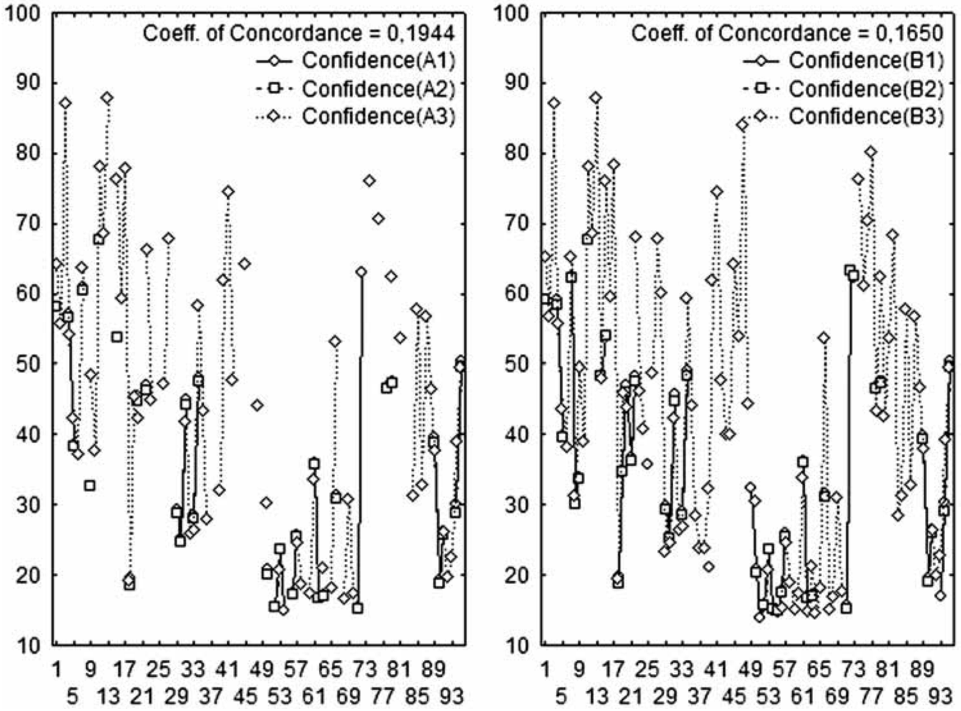
Results are visualized (Fig. 6) in the matrix plot, where an almost perfect concordance in confidence values of the found rules among files (A1, A2, B1, B2) and a smaller, but still significant ($p < 0.05$), concordance among the files with completing paths (A3, B3) and the others was found.

Tab. 7: Kendall Tau correlations for confidence of derived rules

	Con (A1)	Con (A2)	Con (A3)	Con (B1)	Con (B2)	Con (B3)
Con (A1)	1.000000	0.987903	0.731884	0.989305	0.982175	0.763547
Con (A2)	0.987903	1.000000	0.731884	0.975806	0.983871	0.758621
Con (A3)	0.731884	0.731884	1.000000	0.709402	0.720000	0.989071
Con (B1)	0.989305	0.975806	0.709402	1.000000	0.976923	0.754011
Con (B2)	0.982175	0.983871	0.720000	0.976923	1.000000	0.758065
Con (B3)	0.763547	0.758621	0.989071	0.754011	0.758065	1.000000

Source: our own research

Fig. 5: Line plots of multiple variables for confidence of derived rules



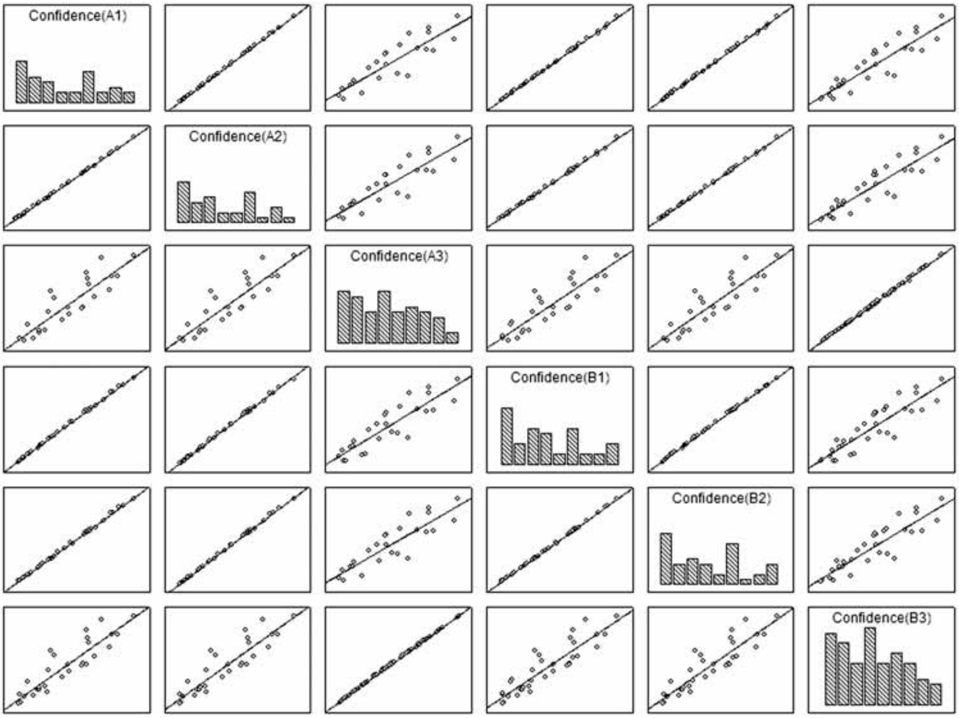
Source: our own research

Data cleaning from crawlers accesses has no impact on the quality of extracted rules in terms of their basic quality characteristics. It has also been proved that completing the paths has a substantial impact on the quality of extracted rules (X3 vs. X1, X2, X3), regardless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY). On the contrary, allowing the used browser upon identifying sessions has not any significant impact on the quality of extracted rules (A1, A2, B1, B2), and in the same way regardless the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY).

Conclusion and Discussion

Web portal of the organization is an information system, which serves as the first place, on which employees, clients and other people look for information connected with, for example, not only offered services, but also tasks and opportunities provided by the organization. For correct functionality of the portal a selection of a suitable system is important, however, more important are information, which appear on the portal [28]. Individual information and their categories, forming mainly items of the menu, are created with the aim to be as transparent, accessible and effective as possible. The order of their arrangement and the contents of individual items are usually created based on the analysis of needs of the portal and classification of information to be provided by the organization. A view of the information can differ depending on the visitor to the portal. This view can differ from the one of the team, which proposed the original structure. Procedures of browsing the information can be understood as patterns of behaviour of users and can be disclosed by means of sequential rules. Thus, we are able to describe, which parts of the portal were visited in the observed period by individual users, and in which order the visits took place. Based on the behaviour patterns of users upon browsing the portal we can execute optimization, i.e. to rearrange, add or remove me-

Fig. 6: Matrix plot for confidence of derived rules



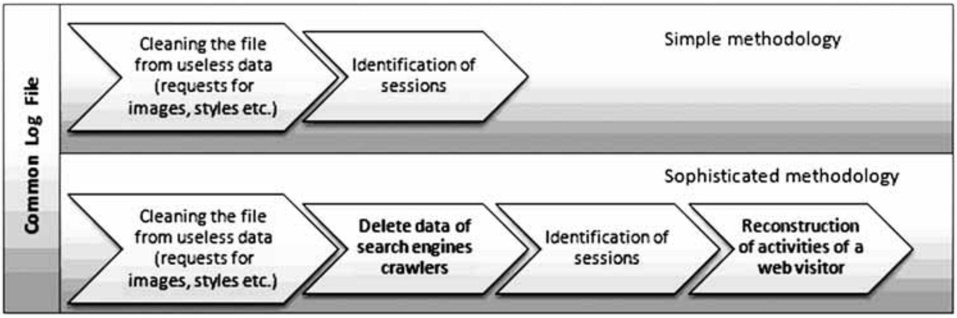
Source: our own research

ssages. By means of removing identified shortcomings we can contribute to the more effective utilization of the portal and to help the portal´s visitors to look the required information up more effectively. Zeithaml et al. [34] developed eleven dimensions of portal qualities: access, ease of navigation, efficiency, flexibility, reliability, personalization, security, responsiveness, assurance/ trust, site aesthetics, and price knowledge. We focus on the ease of navigation and efficiency.

Z. Yang et al. [35] in their research described factors influencing the quality of web portals and considered well organized messages and relevant contents to be the most important factors of usability. Web portals do not represent only a source of information for clients, but also a substantial source of data, from which we can obtain knowledge on visitors of our web portal with the aim to optimize, personalize the portal, analyze the visit rate, etc.. For example, by means of segmentation of users and finding the behaviour patterns of users in identified segments it is possible to adjust individual parts of the portal for different groups. Identification of groups of visitors is important for targeted advertisement on individual sub-pages of the portal, or targeted submission of information for particular groups of users. The knowledge of such segments supports marketing decisions, allows for observing trends in behaviour of users in time, to trace the impact of our decisions on the frequency of visits in individual groups, etc.. By realizing such or similar analyses the organization can find many points about clients and their interests. At present, when web portals of organizations have become an inseparable part of building a brand and communication with customers, such analyses will be more and more necessary and important.

Data on the use of the portal can be obtained by monitoring of the log file of the web server. We are able to create data matrices from accesses and the web map, which will serve for searching for behaviour patterns of users. An advantage of an analysis of the log file of the web

Fig. 7: Proposed methodologies of data preparation based on the experiment results


Common Log File
Simple methodology
Cleaning the file from useless data (requests for images, styles etc.)
Identification of sessions
Sophisticated methodology
Cleaning the file from useless data (requests for images, styles etc.)
Delete data of search engines crawlers
Identification of sessions
Reconstruction of activities of a web visitor

Source: our own research

server comparing to, for example, a selective detection consists in the fact that the visitor does not know that he is the object of investigation. On the other hand, a disadvantage is the mere preparation of data from the log file, which represents the most time-consuming phase of the analysis of the web page. The experiment was realized with the aim to find out to which measure it is necessary to realize this time-consuming preparation of data and we aimed at specifying the steps inevitable for obtaining valid data from the log file. Results of the experiment are very important for the portal which is regularly analysed and adjusted, since they make evidence of the correctness of individual steps, or identify "useless" steps, which can followingly make the data advance preparation simpler.

Summarized results of the realized experiment are as follows:

a. The first assumption concerning data cleaning was proved only partially. Specifically, data cleaning from crawlers accesses has a significant impact on the quantity of extracted rules only in case of files with completing paths (A3 vs. B3). This can be caused by the fact that crawlers of various searching engines browse the web sequentially. If we apply on this searching also paths reconstruction, the programme will generate a number of non-standard data. On the contrary, the impact on the reduction of the portion of inexplicable rules as well as the impact on the quality of extracted rules in terms of their basic quality characteristics was not proved.

b. On the contrary, the second assumption was fully proved – completing paths was found crucial in the data preparation for web log mining. Specifically, it was proved that completing the paths has a significant impact on the quantity of extracted rules (X3 vs. X1, X2, X3). It was similarly proved that path completion has a significant impact on the quality of extracted rules (X3 vs. X1, X2, X3), regardless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY).

c. The third assumption was also fully proved. Specifically, it was proved that allowing the used browser upon identifying sessions has neither significant impact on quantity nor quality of extracted rules (A1, A2, B1, B2), similarly regardless of the fact, whether the files are cleaned from crawlers accesses (BY) or not (AY).

The data themselves are the presumption of each data analysis, regardless its focus (analysis of the visit rate, optimization of the portal, personalization of the portal, etc.). However, results of the analysis depend on the quality of analysed data.

The scheme (Fig. 7) based on the experimental conclusions depicts inevitable steps to be taken for obtaining valid data from the log file. Using the experimental results we propose two alternatives for the log file data processing, while the first one represents a more simple and less time-consuming data preparation at the expense of the lower number and less accurate extracted knowledge. The other, more time-consuming alternative offers more valid data, and thus also a larger amount of better quality knowledge, represented by the found rules. In both provided methodologies of data preparation there absents the step of definition of sessions based on the used web browser, which had neither any impact on quantity or quality of the extracted rules. In the simpler methodology, which represents an alternative without any reconstruction of activities of a web visitor, there also absents the step of cleaning of the file from crawlers of searching services, which had an impact on the quantity of extracted rules only in case of files with the completion of paths.

Searching for frequented rules by means of sequence rule analysis is closely connected with certain steps of data preparation. It was proved that completing the paths is very important, however, it depends to a great degree on the correct identification of individual sessions of the portal visitors. There exist a large number of models for the identification of users sessions [12, 13, 18, 21, 22]. There exists also a method, which expressly identifies these sessions. This method is called additional programming of an application, which creates web logs. We thus obtain a more sophisticated solution, by means of which on the one hand we are able to identify, expressly and confessedly, each visitor´s session, but on the other hand we lose the general method of web log processing. Our next goal is to additionally programme this functionality into our portal and analyse various parameters of individual methods of identification of sessions compared with the reference direct identification.

Path completion also depends on the topicality of the sitemap, which can be modified too quickly so that we could use the offline method of web log analysis. Our further research will be devoted to dynamic analysis of web logs in the real time. Our aim is to reduce time necessary for the advance processing of these logs and at the same time to increase the accuracy of these data depending on the time of their collection.

References

[1] AGRAWAL, R., IMIELINSKI, T., & SWAMI, A. N. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD international conference on management of data*, 1993, Washington, DC, USA, pp. 207–216. ISBN 0-89791-592-5.

[2] BAMSHAD, M. ET AL. Automatic personalization based on Web usage mining. *Communications of the ACM*, 2000, ACM, Vol. 43, No. 8, pp 142-151. ISSN 0001-0782.

[3] BERENDT, B., SPILIOPOULOU, M. Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal,* 2000, Vol. 9, No. 1, pp. 56-75. ISSN 1066-8888.

[4] BERKA, P. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.

[5] BONCHI, F. ET AL. Web log data warehousing and mining for intelligent web caching. D*ata and Knowledge Engineering,* 2001, Vol. 39, No. 2, pp. 165–189. ISSN 0169-023X.

[6] CHEN, Z., SHEN, H. A study of a new method of browsing path data mining. *The sixth international conference of information management research and practice,* 2000, HsingChu, Taiwan, ROC: TsingHua University.

[7] CHEN, M., PARK, J.S, & YU, P.S. Data mining for path traversal patterns in a web environment. *ICDCS*, 1996, pp. 385–392. ISBN 0-8186-7398-2.

[8] COOLEY, R., MOBASHER, B., & SRIVASTAVA, J. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information System,* 1999, Springer-Verlag, Vol. 1, ISSN 0219-1377.

[9] CUNHA, C. R., JACCOUD, C. F. B. Determining WWW user's next access and its application to pre-fetching. *The second IEEE symposium on computers and communications,* 1997, Alexandria, Egypt, pp. 6–11. ISBN 0-8186-7852-6.

[10] ČERNÁ, M., POULOVÁ, P. Návštěvnost portálu a míra využití jejich nástroju a služeb – případová studie. *E+M Ekonomie a Management*, 2008, Vol. 11, No. 4, pp. 132 – 143. ISSN 1212-3609.

[11] HAN, J., PEI, J., & YAN, X. Sequential pattern mining by pattern-growth: Principles and extensions. *Studies in Fuzziness and Soft Computing*, 2005, pp. 180–220.

[12] HAY, B., WETS, G., & VANHOOF, K. Web usage mining by means of multidimensional sequence alignment methods. *WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profile, Lecture Notes in Computer Science,* 2003, Vol. 2703, Springer, Berlin/Heidelberg, pp. 50–65.

[13] HAY, B., WETS, G., & VANHOOF, K. Segmentation of visiting patterns on Web sites using a sequence alignment method. *Journal of Retailing and Consumer Services*, 2003, Vol. 10, No. 3, pp. 145–153. ISSN 0969-6989.

[14] IYENGAR, A., MACNAIR, E., & NGUYEN, T. An analysis of Web server performance. *The IEEE global telecommunications conference,*

1997, Vol. 3, pp. 1943–1947. ISBN 0-7803-4198-8.

[15] JAIN, A. K., MURTY, M. N., & FLYNN, P. J. Data clustering: A review. A*CM Computing Surveys,* 1999, Vol.31, No. 3, pp. 264–323. ISSN 0360-0300.

[16] LIU, B. Web data mining: *Exploring hyperlinks, contents and usage data*. 2007, Springer, ISBN 978-3-540-37881-5.

[17] LOURENÇO, A. G., BELO, O. O. Catching web crawlers in the act. *Proceedings of the 6th international Conference on Web Engineering,* 2006, ICWE '06, Vol. 263, ACM, New York, NY, pp. 265-272. ISBN 1-59593-352-2.

[18] MASSEGLIA, F., TANASA, D., & TROUSSE, B. Webusage mining: Sequential pattern extraction with a very low support. *Advanced Web Technologies and Applications, Lecture Notes in Computer Science,* 2004, Vol. 3007, pp. 513–522. ISBN 978-3-540-21371-0.

[19] MEHTA, M., AGRAWAL, R., & RISSANEN, J. SLIQ - a fast scalable classifier for data mining. *Proceedings of the fifth international conference on extending database technology*, 1996, France, pp. 8–32. ISBN 3-540-61057-X.

[20] MUNK, M. Objavovanie znalostí na základe používania webu – metódy a aplikácie*. Forum Statisticum Slovacum*, 2009, Vol. 5, No. 1, pp. 65-72. ISSN 1336 – 7420.

[21] OYANAGI, S., KUBOTA, K., & NAKASE, A. Mining WWW access sequence by matrix clustering. *Mining Web Data for Discovering Usage Patterns and Profiles, Lecture Notes in Computer Science,* 2003, Vol. 2703, Springer, Berlin/Heidelberg, pp. 119–136. ISBN 978-3540203049.

[22] PARK, S., SURESH, N.C., & JEONG, B. Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering,* 2008, Vol. 65, pp. 512–543. ISSN 0169-023X.

[23] PERKOWITZ, M., ETZIONI, O. Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligence,* 2000, Vol. 118, No. 1–2, pp. 245–275. ISBN 1-55860-709-9.

[24] PIROLLI, P., PITKOW J., & RAO, R. Silk from a sow's ear: Extracting usable structures from the Web. *Proc. of 1996 Conference on Human Factors in Computing Systems (CHI-96)*, 1996, Vancouver.

[25] ROMERO, C., ET AL. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers & Education*, 2009.

[26] SPILIOPOULOU, M., FAULSTICH, L.C. WUM: A Tool for Web Utilization Analysis. *Extended version of Proc. EDBT Workshop WebDB'98,* 1999, Springer Verlag, pp 184–203.

[27] SRIVASTAVA, J., ET AL. Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explorations*, 2000, Vol. 1, pp. 1–12.

[28] SKALKA, J., DRLÍK, M. Modulárny informačný systém univerzitného pracoviska. *UNINFOS 2006: Univerzitné informačné systémy: zborník príspevkov z medzinárodnej konferencie*, 2006, FPV UKF Nitra, pp. 36-41, ISBN 80-8050-976-X.

[29] STANKOVIČOVÁ, I. Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner. *Informační a datová bezpečnost ve vazbě na strategické rozhodování ve znalostní společnosti*, 2009, pp. 1-9. ISBN 978-807318-828-3.

[30] TAUCHER, L., GREENBERG, S. Revisitation patterns in world wide web navigation. *Proc. of Int. Conf. CHI'97*, 1997, Atlanta.

[31] WANG, T., HE, P. Web Log Mining by an Improved AprioriAll Algorithm. *Engineering and Technology,* 2005, No. 4, pp. 97-100.

[32] W3C. Configuration File of W3C httpd [online]. 1995. [cit. 2009-03-10]. Available from <http://www.w3.org/Daemon/User/Config/Logging.html>.

[33] ZAIANE, O., HAN J. WebML: Querzing the World-Wide Web for resources and knowledge. *Workshop on Web Information and Data Management.* 1998, pp. 9-12.

[34] ZEITHAML, V.A., PARASURAMAN, A., & MALHORTA, A. A conceptual framework for understanding e-service quality: implications for future research and managerial practice. *MSI Working Paper Series*, 2001, No. 00-115, Cambridge, MA, pp. 1–49.

[35] YANG, Z., ET AL. Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. *Information & Management*, 2005, Volume 42, Issue 4, pp. 575-589. ISSN 0378-7206.

[36] YU, P. Data mining and personalization technologies. *The sixth IEEE international conference on database systems for advanced applications*, 1999, pp. 6–13. ISBN 0-7695-0084-6.

RNDr. Michal Munk, PhD.
Constantine the Philosopher University in Nitra
Faculty of Natural Sciences
Department of Informatics
mmunk@ukf.sk

PaedDr. Jozef Kapusta, PhD.
Constantine the Philosopher University in Nitra
Faculty of Natural Sciences
Department of Informatics
jkapusta@ukf.sk

PeadDr. Peter Švec
Constantine the Philosopher University in Nitra
Faculty of Natural Sciences
Department of Informatics
psvec@ukf.sk

prof. Ing. Milan Turčáni, CSc.
Constantine the Philosopher University in Nitra
Faculty of Natural Sciences
Department of Informatics
mturcani@ukf.sk

Doručeno redakci: 30. 6. 2009
Recenzováno: 1. 8. 2009, 2. 10. 2009
Schváleno k publikování: 14. 10. 2010

ABSTRACT

DATA ADVANCE PREPARATION FACTORS AFFECTING RESULTS OF SEQUENCE RULE ANALYSIS IN WEB LOG MINING

Michal Munk, Jozef Kapusta, Peter Švec, Milan Turčáni

One of the main tasks of web log mining is discovering patterns of behaviour of portal visitors. Based on the found patterns of users behaviour, which are represented by sequence rules it is possible to modify and improve the web page of an organisation. This article aims at finding out by means of an experiment to what degree it is necessary to realize data preparation for web log mining and it aims also at specifying inevitable steps for obtaining valid data from the log file. Results of the experiment are very important for the portal, which is regularly analysed and modified, since they can prove correctness of individual steps at analysis, or through an identification of "useless" steps they can make the advance preparation of data simpler. These results show that data cleaning from crawlers accesses has a significant impact on the quantity of extracted rules only in case, when we use the method of paths completion. On the contrary, the impact on the reduction of the portion of inexplicable rules as well as the impact on the quality of extracted rules in terms of their basic characteristics was not proved. Paths completing was proved crucial in data preparation for web log mining. It was proved that paths completing has a significant impact both on the quantity and the quality of extracted rules. However, it was proved that allowing the used browser upon identifying sessions has neither any significant impact on the quantity nor on the quality of extracted rules. There exist a number of models for identification of users sessions, which are crucial in data preparation, however, there exists also a method, which identifies them expressly. Our next goal is to additionally programme this functionality into the existing system and analyse various parameters of individual methods of identification of sessions compared with the reference direct identification. It also mentions the necessity to pay attention to the analysis of web logs in the real time and to reduce the time needed for the advance preparation of these logs and at the same time to increase accuracy of these data depending on the time of their collection.

Key Words: *web log mining, data preparation, data quality assessment, sequence rule analysis, patterns, experiment.*

JEL Classification: *C88, L86, M15.*