INFLUENCE OF RATIO OF AUXILIARY PAGES ON THE PRE-PROCESSING PHASE OF WEB USAGE MINING

Michal Munk, Ľubomír Benko, Mikuláš Gangur, Milan Turčáni

Introduction

Business intelligence (BI) uses information technology as a tool for maximizing the competitiveness of businesses. BI allows executives of corporations to better understand the market, their customers and their competitors. Finally, BI helps corporate executives, business managers and other users to make more informed effective strategic decisions. BI encompasses techniques, methodologies, applications and tools for data transformation of raw data into useful information for business analysis purposes. These technologies are able to handle with the large amount of unstructured data and the main goal of BI is to allow for the easy interpretation of these large volumes of data [28]. BI tools include traditional data warehousing technologies like reporting, ad-hoc querying, online analytical processing (OLAP), business performance management, competitive intelligence, benchmarking and predictive analytics.

One of the most useful BI tools is web mining, as the part of data mining (DM). DM is a computational process of the nontrivial extraction of implicit and previously unknown and potentially useful information and patterns in large data sets using methods of artificial intelligence, machine learning, statistics and database systems [9], [11] Web mining is the application of data mining techniques to discover patterns from the web. With respect to the main goal of analysis, web mining can be divided into three types: web content mining, web structure mining and web usage mining. Web content mining is the mining, extraction and integration of useful data and information from web page content. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. The present contribution deals with techniques of web usage mining.

Web usage mining extracts useful information from the web server logs i.e. it is a process of discovering what users are looking for on the examined web site. Large and heterogeneous web logs hide information with great potential value. Web usage mining discovers interesting usage patterns from these logs data in order to better understand web portals users´ behavior and to better serve the needs of web-based applications running on these web portals. Web usage mining is very useful for effective web site management, creating adaptive web sites, business and support services, personalization or network traffic flow analysis. The first step of web usage mining process can be data pre-processing.

Data pre-processing plays an important role in the analysis of user behavior on web portals. From log files we can obtain anonymous information which is needed to be processed before analysis. That is the reason why we need to use methods to identify users based on the log file. In this paper, we will focus on the method called Reference Length, its influence on the ratio of auxiliary pages and its influence on generating useful, trivial and inexplicable rules.

The rest of the paper is structured subsequently: in section 1 we summarize the related work of other authors dealing with business intelligence and data pre-processing. We summarize important data pre-processing methods in section 2. Subsequently, we particularize the experiment in section 3. This section describes the research methodology and provides a summary of the experiment results.

1. Related Work

The web usage mining approach in practical applications of business intelligence has been the objective of many works. In [32],

 DOI: 10.15240/tul/001/2015-3-013

researchers studied customer's behavior using web mining techniques and their application in e-commerce. Data from customers was clustered to segments using the K-Means algorithm, in which input data came from a web log of various e-commerce websites. In [26] authors proposed to use Business Process Management (BPM) methodologies for e-commerce website logs. Web clicks and BPM events were compared, and consequently a methodology for classification and transformation of URLs into events was applied. A general data warehouse/online analytic processing (OLAP) framework for web usage mining and business intelligence reporting was introduced in [13]. In this framework integration of web data warehouse construction, data mining, and OLAP into the e-commerce system dramatically reduced the time and effort for web usage mining, business intelligence reporting, and mining deployment. In [4], data from WiFi hotspots were analyzed to increase coverage and enhance user quality of service with the help of simple K-Means clustering. The recognized patterns in combination with geographic location of WiFi hotspots allowed for making informed decisions including changing customer locations for WiFi hotspots. Various important concepts of web usage mining and its practical application were presented in [1]. The aforementioned solutions comprise the several steps of the web usage mining process. Some of them use data pre-processing techniques as the initial step.

Various methods of data pre-processing were introduced. Researchers in [27] focused on pre-processing techniques implemented on an IIS web server and also proposed some efficient techniques. In [22] a novel pre-processing technique was introduced by removing local and global noise and web robots. Aye in [5] presented two algorithms for field extraction and data cleaning. The researchers in [18] experimented with the accomplishment of pre-processing and clustering of web logs. Kewen [16] introduced algorithms for data pre-processing, that have proved efficient and valid but some related issues need further research.

2. Web Usage Data Pre-Processing

The pre-processing phase is the most important part in any data mining application. Specifically, in web usage mining it is essential to have reliable data from which we can reconstruct user activities on the web portal. For this reason we use log files stored on the web server and record every step the user takes on the web portal. The basic information presented in log files consists of the user name (IP address), visiting path, path traversed, time stamp, page last visited, success rate, user agent, URL and request type. Log files can be located in three different places – on web server, proxy server or on the client side. Mostly they are located on the web server. When a user requests the web site from a particular server, this information will be recorded on the web server [14]. Because of that, the log files often contain a lot of unnecessary information that needs to be filtered. The essential steps of pre-processing of web usage data for knowledge discovery are data cleaning, web user identification, session identification and the reconstruction of activities of a web visitor [7].

2.1 Data Cleaning

Data cleaning is usually site-specific and involves tasks such as removing references to objects that are not important for the reconstruction of user behavior. This includes references to pictures, flash videos, cursors, javascripts or styles. Additionally, data cleaning causes the removal of references due to crawler navigation. Well-known search engine crawlers can be identified and removed thanks to reference in the user-agent field. Other crawlers that avoid recognition, begin their site crawl by first attempting to access to exclusion file "robots.txt" in the server root directory. Therefore these bots can be identified by locating the IP addresses of the accesses to the file [17], [30], [23].

We used this knowledge to create a simple java application (Fig. 1) which helps us with the pre-processing phase of web usage mining and with preparation of the log file for another phase. The algorithm cleans the log file from pictures (*.jpg, *.jpeg, *.png, *.bmp, *.gif), java scripts (*.js), styles (.*css), site summary (*.rss), cursors (*.cur), videos (*.flv, *.swf), favicons (*.ico) and xml files. We also implement the record filter containing HTTP response status codes informing us about errors on the client or the server side. The most important function of our algorithm is the detection of IP addresses which accessed the file "robots.txt". With the use of java class HashSet we contained all of these IP addresses representing crawlers and also all

file types mentioned above as tokens. Then we could use matcher and pattern functions with the java regex library. Matcher [19] is an engine that performs match operations on a character sequence by interpreting a pattern. This way we could find our tokens in lines of the log file. We created a new file for storing the useful data.

2.2 User Identification

The aim of this step is to identify users who visited the web portal. The assumption that IP address alone is enough to identify unique users is incorrect. Behind one IP address can be hidden many more users. For this reason we have to use techniques to identify users such as authentication, cookies or a combination of IP address or the field user agent [7].

If the web portal requests users to register, then the identification of users would be easy. Every unique user name would represent a new user. The access of anonymous users recorded as "-" in the log file would be a problem [7], [24].

If we don't have information about user names, then we could use cookies. When the user requests a web page from the web server, the web server responds with identifying data (cookie) for the user's web browser. By the next visit of the web portal, the web browser will recognize the user. However, this method has problems itself, e.g. anytime a user can manually delete cookies from his/her web browser. For this reason other heuristic methods are used in user identification.

Most of the heuristics use a combination of IP address and another field such as a user's agent. This method has its rules. If the IP address of two records is different, then it is automatically a new user, otherwise we compare the field with the user's agent. If the web browser and operating system are the same in both records, then it is the same user, otherwise it is a new user [7].

Fig. 1: Cleaning the log file

```
public void cleaning
{
  open log file;
  create cleaned log file;
  while not end of file repeat {
    search the line for tokens;
    if match isn't found then
    write line to new file;
    move to another line;
  }
  close both files;
}
```

Source: own

2.3 Session Identification

Every user when browsing the web visits some amount of web pages and spends some time on the web portal. In the process of session identification it is important to divide user's visits into sessions. A session is characterized as the activity of one user in a certain time on the web portal [24].

One solution to the problem of session identification is offered by time oriented heuristics, structure oriented heuristics or navigation oriented user session identification. Cooley et al. [7] introduced time oriented

heuristics called heuristic h1, which creates sessions based on a time window of 30 minutes. Spiliopoulou, Mobasher, Berendt & Nakagawa [29] advised to identify sessions based on a time window of 10 minutes and called it heuristic h2. Structure oriented heuristic h-ref identifies sessions based on the field referrer. If the URL is not followed by the referrer, it becomes a new session [7], [29], [15].

Navigation-oriented methods assume that two sets of transactions, namely auxiliary-content or content-only, can be formed. The reference length approach is based on the assumption that the amount of time a user spends on a page correlates to whether the page should be classified as auxiliary or content page for that user. It is expected that the variance of the time spent on the auxiliary pages is small. If an assumption is made about the percentage of auxiliary references in a log file, a reference length can be calculated, it estimates the cutoff between auxiliary and content references. If we define the assumption about the portion of auxiliary pages in log file, we can define the cutoff time *C*, which separates the content pages from the auxiliary. When the cutoff time is known the session can be created in such a manner that we compare the time of particular web page visit with the cutoff time *C*. The session is then defined as a path through the navigation type of pages (duration of time spent on this web page is less than *C*) to the content page (the user spent there more time than *C*) [7], [15].

We assume that the variance of times spent on the auxiliary pages is small, because the user 'only' passes through the pages to his/her search target. The length of the time spent on content pages has a higher variance.

Provided that the variable *RLength* has an exponential distribution (*Chi-square* = 21.40632; p = 0.06527) and the assumption about the portion of auxiliary pages is created $(0 \leq p < 1)$, we can determine the cutoff time $F^{-1}(p,\lambda) = C = \frac{-1n(1-p)}{\lambda}$ which separates the auxiliary pages from the content ones. The maximum likelihood estimation of the parameter λ is $\hat{\lambda} = \frac{1}{\overline{RLength}}$, where $\overline{RLength}$ is observed mean of times spent on the pages.

If we have an estimation of the cutoff time C, then the visit is a sequence k of the visited pages with the time mark, for which is valid: the first $k-1$ pages are classified as the auxiliary pages, the time spent on these pages is less or equal to the cutoff time and the last k^{th} page is classified as a content one and the time spent on this page is higher than the cutoff time.

Based on the estimation in [15], we created an algorithm for session identification (Fig. 2). First, we converted our cleaned log file to a database where we created a new variable through the stored converted date and time of the visit. Also we should sort the data by IP address. The most important part is the estimation of the cutoff time for which we need the percentage of auxiliary references. We can make a subjective estimate of auxiliary pages based on the web portal, or we can use the sitemap of the web portal. In java we can use a Map library that will help us to store the sitemap and estimate percentage of the auxiliary pages.

2.4 Path Completion

The next step is a reconstruction of activities of a web visitor or path completion. Reconstruction of the activities is focused on a retrograde completion of records on the path the user went through by means of a Back button, since the use of such a button is not automatically recorded into the log file. Having lined up the records according to the IP address we can search for some linkages between the consecutive pages. A sequence for the selected IP address can look like this: A→B→C→D→X. In our example, based on the sitemap the algorithm can find out that there does not exist a hyperlink from page D to our page X. Thus we assume that this page was accessed by the user by means of using a Back button from the one of the previous pages. Then, through backward browsing we find out, on which of the previous pages a reference to page X exists. In our sample case, we can find out that there is no 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 out that there is any hyperlink from page B to page X and so we add page B into the sequence, i.e. A→B→C→D→C→B→X. Finally, our algorithm finds out that page A contains a 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 that the user used the Back button in order to transfer from page D to C, from C to B and from B to A [7], [20].

Fig. 2: Session identification

```
public void rlength (double p)
{
  C = estimate of the cutoff time (p);
  while not end of list repeat {
    if ip address of two following records is the same
     then if difference of their unixtime > C
                then new session
                else same session;
    else new session;
  }
}
```

Source: own

3. Experiment

For the comparison of the methods calculating the cutoff time *C* we used the log file of a university web site. Records were cleaned using conventional log file pre-processing methods. Redundant data such as pictures, javascripts, cursors and cascade styles as well as records about robots were removed through our algorithm. Next steps involved user identification and session identification which can be achieved differently. Regarding the evaluation of the pre-processing phase we decided to disregard the user identification and focus on session identification using the Reference Length technique and also on the reconstruction of the activities of web users using the sitemap. In the experiment, we compared four files which were prepared on various levels. Each file was cleaned of unnecessary data with the same algorithm. In the phase of session identification we used the Reference Length method with the difference of calculation cutoff time. We compared the influence of the ratio of auxiliary pages on the calculation based on the sitemap and subjective estimation. Included in the subjective estimate is the decision that the page is auxiliary, based on what the creator or administrator of the web portal defines as an auxiliary page. Typically, in subjective estimation the ratio of the auxiliary page is used. On the other hand the alternative could be provided by the ratio calculated from the sitemap. In this approach we defined every page that has a subpage as an auxiliary page.

3.1 Research Methodology

The experiment was realized in several steps as in [20], [21].

1. Data acquisition – defining the observed variable 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 of the web content).
3. Data preparation on various levels:
   - 3.1 with an identification of sessions – Reference Length calculated from the sitemap (File A1),
   - 3.2 with an identification of sessions – Reference Length calculated from the sitemap and completing the paths (File A2),
   - 3.3 with an identification of sessions – Reference Length calculated from subjective estimate (File B1),
   - 3.4 with an identification of sessions – Reference Length calculated from

subjective estimate and completing the paths (File B2).
4. Data analysis – searching for behavioral 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 elaborates on various levels of data preparation from the point of view of quantity and quality of the found rules.

We articulated the following assumptions:
1. We expect that the identification of sessions by Reference Length, calculated from the sitemap, will have a significant impact on the quantity of extracted rules in terms of decreasing the portion of trivial and inexplicable rules.
2. We expect that the identification of sessions by Reference Length, calculated from the sitemap, will have a significant impact on the quality of extracted rules in the term of their basic measures of quality.

In File A1, sessions were identified using the Reference Length method and the ratio of auxiliary pages was calculated from the sitemap (12.3%). In File A2, sessions were identified using the Reference Length, the ratio of auxiliary pages was calculated from the sitemap (12.3%) and we reconstructed the activities of web users. In File B1, sessions were identified using the Reference Length and the ratio of auxiliary pages was a subjective estimate (30%). In File B2, sessions were identified using the Reference Length and the ratio of auxiliary pages was a subjective estimate (30%) and we reconstructed the activities of web users. In the next steps we applied sequence rule analysis to these files to extract sequence rules for each file. Finally, we joined these rules into one data matrix where each rule can occur once.

3.2 Results

After data cleaning and sequence (session) identification using the Reference Length method we obtained 154,681 numbers of accesses to web portal in both file A1 and B1. After the reconstruction of activities of web users we obtained 178,043 accesses in file A2. In comparison to file B2 we recorded an increase to 178,806 accesses but this is not a relevant difference. Path completion does not have an influence on sequence identification. On the other hand, the ratio of auxiliary pages has a significant influence on sequence identification. In files A1 and A2 we discovered 51,098 sequences and in files B1 and B2 we discovered 20% less sequences. The number of frequent sequences is different in each file. Table 1 depicts the number of accesses, sessions as well as the number of extracted rules.

Using the sequence rule analyses we extracted sequence rules from the frequent sequences with the minimum support 0.01 for each file. By joining the rules to a single data matrix we got 78 unique rules. From this we identified 35 (45%) rules in file A1, 39 (50%) rules in file B1, 62 (79%) in file A2 and 77 (99%) in file B2.

Tab. 1: Number of accesses, sequences and rules

	File A1	File A2	File B1	File B2
Number of accesses	154,681	178,043	154,681	178,806
Number of identified sequences (sessions)	51,098	51,098	40,756	40,756
Number of frequent sequences	37	57	43	70
Absolute number of extracted rules	35	62	39	77
Relative number of extracted rules	0.45	0.79	0.50	0.99
Number of actionable rules	10	10	10	10
Number of trivial rules	17	35	19	40
Number of inexplicable rules	8	17	10	27

Source: own

With sequence rule analysis we can get actionable (useful), trivial and inexplicable rules. To decide on the type of rule, there is no algorithm. Useful rules contain high quality, actionable information. Trivial results are already known by anyone at all familiar with the business. Inexplicable results seem to have no explanation and do not suggest a course of action [6]. In our research we found similar number of actionable rules in every file. In files with session identification by Reference Length, A1 and B1, we found 17 and 19 trivial rules. By using path completion in files A2 and B2, we identified 35 and 40 trivial rules. The greatest difference in rules identification was found by inexplicable rules, where the number of rules was three times higher in files A2 and B2 than in files where we did not use path completion.

Comparison of the Portion of Found Rules in Examined Files

The analysis (Tab. 2) resulted in sequence rules which we obtained from frequent sequences fulfilling their minimum support (in our case, min *s* = 0.01). Frequent sequences were obtained from identified sequences, i.e. visits of individual users during one week. We used STATISTICA Sequence, Association and Link Analysis, for sequence rules extraction. It is an implementation of algorithm using the powerful a-priori algorithm [2], [3], [10], [31] together with a tree structured procedure that only requires one pass through the data [8].

There is a high coincidence between the results (Tab. 2) of sequence rule analysis in terms of the portion of found rules in the case of files with the identification of sessions based on sitemap estimation and subjective estimation without path completion (A1, B1). The most rules were extracted from files with path completion; concretely 62 were extracted from the file A2, which represents over 79% and 77 were extracted from the file B2, which represents over 98% of the total number of found rules. Generally, more rules were found in the observed files with the completion of the paths.

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 usage mining, is rejected at the 1% significance level.

The following graph (Fig. 3) visualizes the results of Cochran´s Q test.

Kendall´s coefficient of concordance represents the degree of concordance in the number of found rules among examined files. The value of coefficient (Tab. 3) is approximately 0.37, while 1 means a perfect concordance and 0 represents a discordance. Low values of coefficient confirm the Q test results.

From multiple comparisons [25] one homogenous group (Tab. 3) was identified in

Tab. 2: Incidence of discovered sequence rules in particular files

Body	=>	Head	A1	A2	B1	B2	Type of rule
(http://www.ukf.sk)	=>	(http://www.ukf.sk/o-univerzite/adresar)	1	1	1	1	useful
⋮	=>	⋮	⋮	⋮	⋮	⋮	⋮
(http://www.ukf.sk/struktura-univerzity), (http://www.ukf.sk)	=>	(http://www.ukf.sk/struktura-univerzity/filozoficka-fakulta)	0	0	0	1	Inexplicable
⋮	=>	⋮	⋮	⋮	⋮	⋮	⋮
(http://www.ukf.sk)	=>	(http://www.ukf.sk/prijimaciekonanie)	1	1	1	1	trivial
Count of derived sequence rules			35	62	39	77	
Percent of derived sequence rules (Percent 1's)			44.9	79.5	50.0	98.7	
Percent 0's			55.1	20.5	50.0	1.3	
Cochran Q test			Q = 86.63190; df = 3; p < 0.001				

Source: own

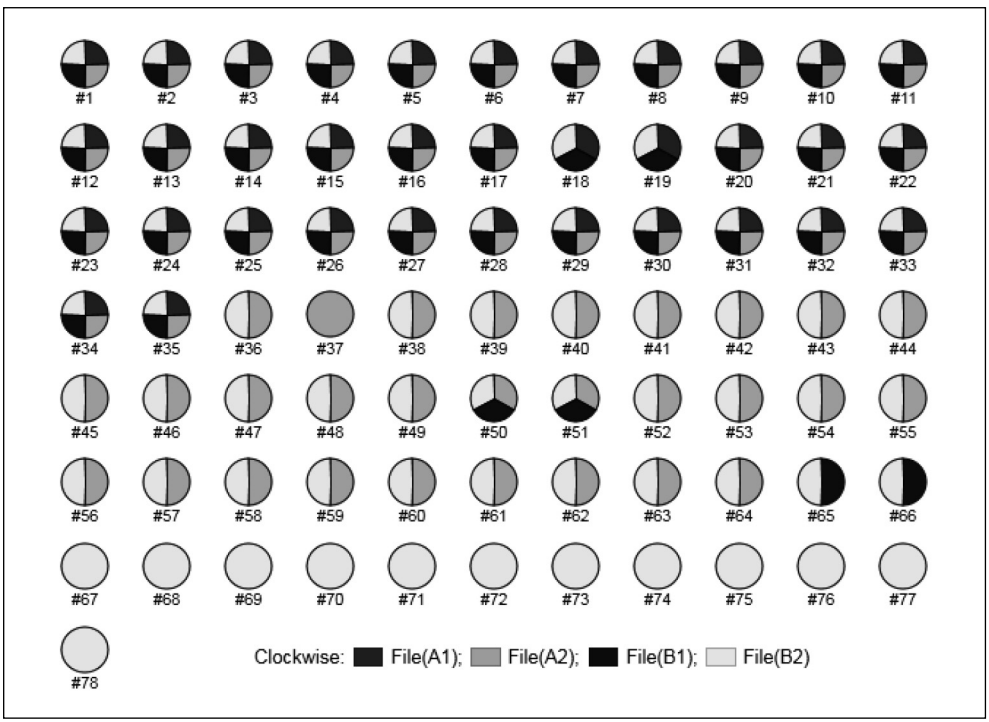
terms of the average incidence of found rules (A1, B1). Statistically significant differences were proved on the level of significance 0.05 in the average incidence of found rules between files A2 and B2 as well as X2 and X1.

The ratio of auxiliary pages has an important impact on the quantity of extracted rules only in the case of path completion (A2, B2).

If we look closely at the results (Tab. 4), we can see that in files without completion of the paths (A1, B1) identical rules were found, except four rules in the case of file with subjective estimation of the ratio of auxiliary pages (B1).

On the other hand (Tab. 5), a statistically significant difference was proved in the case of files with the completion of paths (A2, B2). The

Fig. 3: Icon plot for derived rules in examined files



Source: own

difference consisted of 16 new rules which were found in the file with subjective estimation of the ratio of auxiliary pages (B2).

In the case of files without path completion (A1, B1) the portion of new files represented 5% (Tab. 4). In the case of files with path completion (A2, B2) it is almost 21%, where also the statistically significant difference in the number of found rules between A2 and B2 in favor of B2 was proved (Tab. 5).

Comparison of the Portion of Inexplicable Rules in Examined Files

Now, we will look at the results of sequence rule analysis more closely, taking into consideration the portion of each kind of the discovered rules. From the association rules, we require rules which are not only clear but also useful. Association analysis produces the three common types of rules [6]:

- useful (utilizable, beneficial),

Tab. 3: Homogeneous groups for incidence of derived rules in examined files

File	Incidence Mean	1	2	3
A1	0.449	***		
B1	0.500	***		
A2	0.795		***	
B2	0.987			***
Kendall Coefficient of Concordance		0.37022		

Source: own

Tab. 4: Crosstabulations: File A1 x File B1

A1\B1	0	1	Σ
0	39	4	43
	50.00%	5.13%	55.13%
1	0	35	35
	0.00%	44.87%	44.87%
Σ	39	39	78
	50.00%	50.00%	100.00%
McNemar (B/C)	Chi-square = 2.25000; df = 1; p = 0.134		

Source: own

Tab. 5: Crosstabulations: File A2 x File B2

A2\B2	0	1	Σ
0	0	16	16
	0.00%	20.51%	20.51%
1	1	61	62
	1.28%	78.21%	79.49%
Σ	1	77	78
	1.28%	98.72%	100.00%
McNemar (B/C)	Chi-square = 11.52941; df = 1; p = 0.000694		

Source: own

- trivial,
- inexplicable.

In our case, upon sequence rules, we will differentiate the same types of rules. The only requirement (validity assumption) of the use of chi-square test is high enough expected frequencies [12]. The condition is violated if the expected frequencies are lower than 5. The validity assumption of chi-square test is violated in our tests. This is the reason why we shall not prop ourselves only upon the results of Pearson

chi-square test, but also upon the value of the calculated contingency coefficient.

Contingency coefficients (Coef. C, Cramér's V) represent the degree of dependency between two nominal variables. The value of coefficient (Tab. 6) is approximately 0.40. There is a medium dependency among the portion of useful, trivial and inexplicable rules and their occurrence in the set of discovered rules extracted from the file A1, the contingency coefficient is statistically significant. The zero hypothesis (Tab. 6) is rejected at the 1% significance level, i.e. the portion of useful, trivial and inexplicable rules depends on the identification of sessions based on sitemap estimation. In this file, the least trivial and inexplicable rules were found, while 10 useful rules were extracted from the file A1 which represents 100% of all the found useful rules.

The value of coefficient (Tab. 7) is approximately 0.37, where 1 means perfect relationship and 0 no relationship. There is a medium dependency among the portion of useful, trivial and inexplicable rules and their occurrence in the set of the discovered rules extracted from the file B1, the contingency coefficient is statistically significant. The zero hypothesis (Tab. 7) is rejected at the 1% significance level,

Tab. 6: Crosstabulations: Incidence of rules x Types of rules: File A1

A1\Type	useful	trivial	inexplicable
0	0	24	19
	0.00%	58.54%	70.37%
1	10	17	8
	100.00%	41.46%	29.63%
Σ	10	41	27
	100%	100%	100%
Pearson	Chi-square = 15.01403; df = 2; p = 0.001		
Con. Coef. C	0.40177		
Cramér's V	0.43783		

Source: own

Tab. 7: Crosstabulations: Incidence of rules x Types of rules: File B1

B1\Type	useful	trivial	inexplicable
0	0	22	17
	0.00%	53.66%	62.96%
1	10	19	10
	100.00%	46.34%	37.04%
Σ	10	41	27
	100%	100%	100%
Pearson	Chi-square = 12.03433; df = 2; p = 0.002		
Con. Coef. C	0.36560		
Cramér's V	0.39279		

Source: own

Tab. 8: Crosstabulations: Incidence of rules x Types of rules: File A2

A2\Type	useful	trivial	inexplicable
0	0	6	10
	0.00%	14.63%	37.04%
1	10	35	17
	100.00%	85.37%	62.96%
Σ	10	41	27
	100%	100%	100%
Pearson	Chi-square = 7.97115; df = 2; p = 0.019		
Con. Coef. C	0.30450		
Cramér's V	0.31968		

Source: own

Tab. 9: Crosstabulations: Incidence of rules x Types of rules: File B2

B2\Type	useful	trivial	inexplicable
0	0	1	0
	0.00%	2.44%	0.00%
1	10	40	27
	100.00%	97.56%	100.00%
Σ	10	41	27
	100%	100%	100%
Pearson	Chi-square =0.91416; df = 2; p = 0.633		
Con. Coef. C	0.10763		
Cramér's V	0.10826		

Source: own

i.e. the portion of useful, trivial and inexplicable rules depends on the identification of sessions based on subjective estimation.

The value of coefficient (Tab. 8) is approximately 0.30, where 1 means perfect relationship and 0 no relationship. There is a medium dependency among the portion of useful, trivial and inexplicable rules and their occurrence in the set of discovered rules extracted from the file A2, the contingency coefficient is statistically significant. The zero hypothesis (Tab. 8) is rejected at the 5% significance level, i.e. the portion of useful, trivial and inexplicable rules depends on the identification of sessions based on sitemap estimation and path completion.

The coefficient value (Tab. 9) is approximately 0.11, where 1 represents perfect dependency and 0 means independency. There is a little dependency among the portion of useful, trivial and inexplicable rules and their occurrence in the set of discovered rules extracted from the file B2, and the contingency coefficient is not statistically significant. In this file, the most trivial and inexplicable rules were found, while the portion of useful rules has not changed.

This corresponds with results from the previous chapter *Comparison of the portion*

Tab. 10: **Homogeneous groups for (a) support of derived rules; (b) confidence of derived rules**

(a)				
File	**Support Mean**	**1**	**2**	**3**
A1	3.425	****		
B1	3.941	****	****	
A2	4.163		****	
B2	4.747			****
Kendall Coefficient of Concordance		0.63692		

(b)			
File	**Confidence Mean**	**1**	**2**
A1	15.942		****
B1	17.123	****	****
A2	22.320	****	
B2	23.442	****	
Kendall Coefficient of Concordance		0.49568	

Source: own

of found rules in examined files, where the significant differences in the number of discovered rules between files A1, B1 was not proved. On the contrary, there was a statistically significant difference between A2 and B2 in favor of B2. If we look at the differences between A2 and B2 in dependency on types of rule (Tab. 8, Tab. 9), we observe an increase in the number of trivial and inexplicable rules in case B2, while the portion of useful rules is equal in both files.

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

Quality of sequence rules is assessed by means of two indicators [6]:

- support,
- confidence.

The results of the sequence rule analysis showed differences not only in the quantity of the found rules but also in the quality. Kendall´s coefficient of concordance represents the degree of concordance in the support of the found rules among examined files. The value of coefficient (Tab. 10a) is approximately 0.64, while 1 means a perfect concordance and 0 represents discordancy.

From the multiple comparisons (Scheffe test) two homogenous groups (Tab. 10a) consisting of examined files were identified in terms of the average support of the found rules. The first homogenous group consists of files A1, B1 and the second of files B1, A2. There is not a statistically significant difference in support of discovered rules between these files. On the contrary, statistically significant differences on the level of significance 0.05 in the average support of found rules were proved among files A1, A2, B2 and between files B1, B2.

There were demonstrated differences in the quality in terms of confidence characteristics values of the discovered rules among individual files. The coefficient of concordance values (Tab. 10b) is almost 0.50, while 1 means a perfect concordance and 0 represents discordancy.

From the multiple comparisons (Scheffe test) two homogenous groups (Tab. 10b) consisting of examined files were identified in term of the average confidence of the found rules. The first homogenous group consists of files B1, A2, B2 and the second of files A1, B1. There is not a statistically significant difference in confidence of discovered rules between these files On the contrary; statistically significant

differences on the level of significance 0.05 in the average confidence of found rules were proved between files A1, A2 and between files A1, B2.

Conclusions

This paper was intended to compare the influence of the ratio of auxiliary pages on the calculation of cutoff time in the reference length method. Both assumptions concerning the quantity and the quality of extracted rules of sessions identified using the Reference Length method, calculated from the sitemap were only proven partially. They were fully proven after path completion. On the contrary, path completion is dependent on the accuracy of session identification.

The ratio of auxiliary pages has the impact on the quantity of extracted rules only in the case of files with path completion (A2 vs. B2). However, making provisions for the identification of sessions based on the estimation of the ratio of auxiliary pages has no significant impact on the quantity of extracted rules in the case of files without path completion (A1 vs. A2).

The portion of trivial and inexplicable rules is dependent on the estimation of the ratio of auxiliary pages by the sessions´ identification based on Reference Length in the case of the reconstruction of a user`s activities. Session identification based on the sitemap has no impact on increasing number of useful rules. On the contrary, inappropriate estimation of the ratio of auxiliary pages may cause an increasing number of trivial and inexplicable rules.

Results show that the largest degree of concordance in support and confidence is among the rules found in files without path completion (A1, B1). On the contrary, discordancy in support is between files with various estimations of the ratio of auxiliary pages in case of path completion (A2, B2). Estimation of the ratio of auxiliary pages by identification of sessions based on Reference Length has a substantial impact on the quality of extracted rules in case of the reconstruction of user`s activities.

The Reference Length method is a good option for session identification. The disadvantage of using the Reference Length is the need for exponential distribution of the variable *RLength* that has to be examined otherwise it cannot be used for session identification. Different approaches to estimation of the ratio of auxiliary pages have been shown to have an impact after path completion. It is recommended to use the calculation of ratio based on the sitemap because it is more accurate than a subjective estimation. The disadvantage of sitemap is that the web portal undergoes frequent changes and the sitemap used for the calculation could be different with time.

Future work may involve the optimization of our proposed algorithms and creating an algorithm to automatically return the ratio of auxiliary pages from the sitemap.

The results show the importance of the reconstruction of user's activities to follow the accurate estimation of the ratio of auxiliary pages. They show the impact of this estimation on the quantity and also on the quality of extracted rules. A sufficient number of quality rules allows sophisticated analysis of the user's behavior on the web site. These analyses results support executives to make effective decisions in future web site customization and in setting strategy for future marketing campaigns. In this way the proposed algorithms increase the usefulness and the importance of described web usage mining technique as a tool of business intelligence.

This paper is supported by the project VEGA 1/0392/13 Modelling of Stakeholders' Behaviour in Commercial Bank during the Recent Financial Crisis and Expectations of Basel Regulations under Pillar 3- Market Discipline.

References

[1] ABRAHAM, A. Natural computation for business intelligence from Web usage mining. In: *Proceedings of Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. 2005, pp. 3-10. DOI: 10.1109/SYNASC.2005.59.

[2] AGRAWAL, R., IMIELIŃSKI, T., SWAMI, A. Mining Association Rules Between Sets Of Items In Large Databases. In: *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data.* New York: ACM, 1993, pp. 207-216. ISBN 0-89791-592-5. DOI: 10.1145/170036.170072.

[3] AGRAWAL, R., SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases.* San Francisco, CA: Morgan Kaufmann Publishers Inc., 1994. pp. 487-499.

[4] ARORA, D., NEVILLE, S.W., LI, K.F. Mining WiFi Data for Business Intelligence. In: *8th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC).* IEEE, 2013. pp. 394-398. DOI: 10.1109/3PGCIC.2013.67.

[5] AYE, T. Web log cleaning for mining of web usage patterns. In: *Computer Research and Development (ICCRD)*. Vol. 2. IEEE, 2011. pp. 490-494. ISBN 978-1-61284-839-6. DOI: 10.1109/ICCRD.2011.5764181.

[6] BERRY, M., LINOFF, G. *Data mining techniques for marketing, sales, and customer relationship management.* 2nd ed. Indianapolis: Wiley, 2004. 672 p. ISBN 978-0-471-47064-9.

[7] COOLEY, R., MOBASHER, B., SRIVASTAVA, J. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems.* 1999, Vol. 1, Iss. 1, pp. 5-32. ISSN 0219-1377. DOI: 10.1007/BF03325089.

[8] *Electronic statistics textbook*. Tulsa, OK: Statsoft, 2010.

[9] FRAWLEY, W., PIATETSKY-SHAPIRO, G., MATHEUS, C. Knowledge Discovery in Databases: An Overview. *AI Magazine*. 1992, Vol. 13, Iss. 3, pp. 213-228. ISSN 0738-4602. DOI: 10.1609/aimag.v13i3.1011.

[10] HAN, J., LAKSHMANAN, L., PEI, J. Scalable Frequent-pattern Mining Methods: An Overview. In: *Tutorial Notes of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2001. pp. 5.1-5.61. DOI: 10.1145/502786.502792.

[11] HAND, D., MANNILA, H., SMYTH, P. *Principles of Data Mining*. MIT Press, 2001. 584 pp. ISBN 978-0262082907.

[12] HAYS, W. *Statistics*. 4th ed. New York: CBS College Publishing, 1988. 750 p. ISBN 978-0030024641.

[13] HU, X.H., CERCONE, N. A data warehouse/online analytic processing framework for web usage mining and business intelligence reporting. *International Journal of Intelligent Systems*. 2004, Vol. 19, Iss. 7, pp. 585-606. ISSN 1098-111X. DOI: 10.1002/int.20012.

[14] JOSHILA GRACE, L., MAHESWARI, V., NAGAMALAI, D. Web Log Data Analysis and Mining. *Advanced Computing*. 2011, Vol. 133, pp. 459-469. ISSN 1865-0929. DOI: 10.1007/978-3-642-17881-8_44.

[15] KAPUSTA, J., MUNK, M., DRLIK, M. Cut-off time calculation for user session identification by reference length. In: *Application of Information and Communication Technologies.* IEEE, 2012. pp. 1-6. ISBN 978-1-4673-1739-9. DOI: 10.1109/ICAICT.2012.6398500.

[16] KEWEN, L. Analysis of preprocessing methods for web usage data. In: *Measurement, Information and Control (MIC)*. Vol. 1. IEEE, 2012. pp. 383-386. ISBN 978-1-4577-1601-0. DOI: 10.1109/MIC.2012.6273276.

[17] LIU, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. 2nd ed. Berlin: Springer, 2011. 624 p. ISBN 978-3-642-19459-7.

[18] MAHESWARI, B., SUMATHI, P. A New Clustering and Preprocessing for Web Log Mining. In: *Computing and Communication Technologies (WCCCT).* IEEE, 2014. pp. 25-29. ISBN 978-1-4799-2876-7. DOI: 10.1109/WCCCT.2014.67.

[19] *Matcher (Java Platform SE 7)* [online]. 2014 [cit. 2015-05-10]. Available from: http://docs.oracle.com/javase/7/docs/api/java/util/regex/Matcher.html.

[20] MUNK, M., KAPUSTA, J., ŠVEC, P. Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor. *Procedia Computer Science*. 2010, Vol. 1, Iss. 1, pp. 2273-2280. ISSN 1877-0509. DOI:10.1016/j.procs.2010.04.255.

[21] MUNK, M., KAPUSTA, J., ŠVEC, P., TURČÁNI, M. Data advance preparation factors affecting results of sequence rule analysis in Web Log Mining. *E+M Ekonomie a Management.* 2010, Vol. 13, Iss. 4, pp. 143-160. ISSN 1212-3609.

[22] NITHYA, P., SUMATHI, P. Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise, Cookies and Web Robots. *International Journal of Computer Applications.* 2012, Vol. 53, Iss. 17, pp. 1-6. ISSN 0975-8887. DOI: 10.5120/8510-1684.

[23] PAMUTHA, T., CHIMPHLEE, S., KIMPAN, C., SANGUANSAT, P. Data preprocessing on Web Server Log Files for Mining User Access Patterns. *International Journal of Research and Reviews in Wireless Communications*. 2012, Vol. 2, Iss. 2, pp. 92-98. ISSN 2046-6447. Available also from: http://sci-tech.dusit.ac.th/page/research/siriporn.pdf.

[24] PATIL, P., PATIL, U. Preprocessing of web server log file for web mining. *World Journal of Science and Technology*. 2012, Vol. 2, Iss. 3, pp. 14-18. ISSN 2231-2587.

[25] PILKOVA, A., VOLNA, J., PAPULA, J., HOLIENKA, M. The Influence of Intellectual Capital on Firm Performance Among Slovak

SMEs. In: *Proceedings of the 10th International Conference on Intellectual Capital, Knowledge Management and Organisational Learning (ICICKM-2013)*, Reading: Academic Conferences and Publishing International Limited, 2013. pp. 329-338. ISBN 978-1-909507-80-7.
[26] POGGI, N., MUTHUSAMY, V., CARRERA, D., KHALAF, R. Business process mining from e-commerce web logs. *Lecture Notes in Computer Science*. 2013, Vol. 8094, pp. 65-80. ISSN 0302-9743. DOI: 10.1007/978-3-642-40176-3_7.
[27] REDDY, K., VARMA, G., BABU, I. Preprocessing the web server logs: an illustrative approach for effective usage mining. *ACM SIGSOFT Software Engineering Notes.* 2012, Vol. 37, Iss. 3, pp. 1-5. DOI: 10.1145/180921.2180940.
[28] RUD, O. *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. Hoboken, NJ: Wiley & Sons, 2009. 283 p. ISBN 978-0-470-39240-9.
[29] SPILIOPOULOU, M., MOBASHER, B., BERENDT, B., NAKAGAWA, M. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS Journal on Computing.* 2003, Vol. 15, Iss. 2, pp. 171-190. ISBN 1526-5528. DOI: 10.1287/ijoc.15.2.171.14445.
[30] SUMATHI, C., PADMAJA VALLI, R., SANTHANAM, T. An overview of preprocessing of web log files for web usage mining. *Journal of Theoretical and Applied Information Technology*. 2011, Vol. 34, Iss. 1, pp. 88-95. ISSN 1992-8645.
[31] WITTEN, I., FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques.* 1st ed. Morgan Kaufmann Publishers Inc., 2000. ISBN 978-1558605527.
[32] YADAV, M.P., FEEROZ, M., YADAV, V.K. Mining the customer behavior using web usage mining in e-commerce. In: *Third international conference on computing communication & networking technologies (ICCCNT).* IEEE, 2012. pp. 1-5. DOI: 10.1109/ICCCNT.2012.6395938.

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

Mgr. Ľubomír Benko
University of Pardubice
Faculty of Economics and Administration
Institute of System Engineering and Informatics
Constantine the Philosopher University in Nitra
Faculty of Natural Sciences
Department of Informatics
lubomir.benko@gmail.com

RNDr. Mikuláš Gangur, Ph.D.
University of West Bohemia in Pilsen
Faculty of Economics
Department of Economics and Quantitative Methods
gangur@kem.zcu.cz

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

Abstract

INFLUENCE OF RATIO OF AUXILIARY PAGES ON THE PRE-PROCESSING PHASE OF WEB USAGE MINING

Michal Munk, Ľubomír Benko, Mikuláš Gangur, Milan Turčáni

Data mining belongs to the one of the important tools for Business Intelligence. It is a means to increase competitiveness of a company. Web usage mining is engaged in data mining of web server log file and it analyzes the user´s behavior on the web site. The first step of web usage mining process is data pre-processing obtained from a web log file. Data pre-processing is an important part of web usage mining. Discovering patterns of behavior of web visitors depends on the quality of pre-processing phase. Therefore it is important to understand the used methods. This paper summarizes the pre-processing phases and especially the phases of session identification. There are introduced two algorithms for data cleaning and session identification using the reference length method. The main aim of this paper is to compare a calculation of cutoff time and its influence on discovered useful, trivial and inexplicable rules. Cutoff time is an important part of the session identification using the Reference Length method. The influence of ratio of auxiliary pages on the calculation based on a sitemap and subjective estimation was compared. Statistical methods were used to determine the difference between these two approaches. In this paper was examined the portion of found rules based on quantity and quality. The ratio of auxiliary pages has only an impact on quantity of extracted rules in the files with path completion. It has no impact on portion of extracted useful rules, on the other hand, inappropriate estimation of the ratio of auxiliary pages may cause increasing of trivial and inexplicable rules.

Key Words: *Web usage mining, data pre-processing, session identification, auxiliary pages, reference length, log files, business intelligence, data mining.*

JEL Classification: *C88, C69, M15, O33, D89.*

DOI: *10.15240/tul/001/2015-3-013*