Procedia Computer Science

Volume 29, 2014, Pages 1779–1790

ICCS 2014. 14th International Conference on Computational Science

Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal

Jozef Kapusta[1], Michal Munk[1], Peter Svec[1*] and Anna Pilkova[2]

[1]*Constantine the Philosopher University in Nitra, Nitra, Slovakia*

[2]*Commenius University in Bratislava, Bratislava, Slovakia*

{jkapusta,mmunk,psvec}@ukf.sk, anna.pilkova@fm.uniba.sk

Abstract

In this paper, we focus on finding the suitable value of the time threshold, which is then used in the method of user session identification based on the time. To determine its value, we used the Length variable representing the time a user spent on a particular site. We compared two values of time threshold with experimental methods of user session identification based on the structure of the web: Reference Length and H-ref. When comparing the usefulness of extracted rules using all four methods, we proved that the use of the time threshold calculated from the quartile range is the most appropriate method for identifying sessions for web usage mining.

Keywords: session time threshold, session identification, web log mining, data preprocessing methodology, market discipline

1 Introduction

Web servers offer a valuable source of information through the hosted websites. Web servers also keep information about visitors. This information can be later used for the analysis of visitor behaviour. While the primary web data are used to get knowledge from the web structure or its content, the secondary ones are used in the WUM. The source of secondary data can be a web server access or error log file, proxy server log file, web browser log, browser cookies etc. A web server log is in its default form known as Common Log File and keeps information about IP address; date and time of visit; accessed and referenced resource. If we use the extended version of the log file, we can collect more information, e.g. type of browser (User-Agent Field). In the first phase, web log data are pre-processed in order to identify users, sessions, page views, and clickstreams. Pre-processing refers

* Corresponding author: Constantine the Philosopher University in Nitra, Department of Computer Science, Tr. A.Hlinku 1, 949 74 Nitra, Slovakia; email:psvec@ukf.sk

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014

doi:10.1016/j.procs.2014.05.163

to the stage of processing the web server logs to identify meaningful representations. Data cleaning methods are necessary because web usage mining is sensitive to noise. On the other hand, data pre-processing can be a difficult task when the available data is incomplete or includes erroneous information. According to Cooley et al. (Cooley, Mobasher, Srivastava, & others, 1999) it consists of

1. data cleaning (for removing irrelevant references and fields, removing erroneous references, adding missing references due to caching mechanisms, etc.),
2. data integration (for synchronizing data from multiple server logs, integrating registration data, etc.)(Bayir, Toroslu, & Cosar, 2006),
3. data transformation (for user-session identification (Chen, Fu, & Tong, 2003), path completion (Li, Feng, & Mao, 2008; C. Zhang & Zhuang, 2008),
4. data reduction (for reducing dimensionality) (Nasraoui & Saka, 2007).

Pattern discovery of web usage mining which is the outcome of the proposed methodology is discussed in detail in (Klocoková, 2011; Koprda, Balogh, & Turčáni, 2011; Nina, Rahman, Bhuiyan, & Ahmed, 2009; Song, Luo, Chen, & Gao, 2007; Tan, Chen, & Yang, 2010; Turcinek, Stastny, & Motycka, 2012). In general, discovering association and sequence rules, segmentation (cluster analysis, methods based on analogy, etc.), and classification (decision rules, decision trees, Bayes classification, etc.) are the most applied methods in the web usage mining. On the contrary, a prediction (multinomial logit model, neural networks, support vector regression, etc.) is less common in the web usage mining (Munk & Drlík, 2014; Olej & Filipova, 2012; Olej & Filipová, 2011).

In this paper, we focus on user session identification that is the third step of data pre-processing as stated above. The easiest way to identify the session is to split visits based on the visitor IP address, but then the problem of the unique IP address raise. If the user's network use the NAT (or proxy) technology, which is actually common in most networks, more than one user is hidden under one public address. Another problem the identification of unique visit raises when the user use more than one computer (multiple IP addresses) at the same time, or she uses more browsers at the same computer.

The most commonly used technique for user session identification is the time threshold (Berendt & Spiliopoulou, 2000). If we find two consecutive records with a time difference higher than the selected time threshold, we will split the session. This method is simple to implement with multiple threshold values, so it is very popular.

We can get another view for the session identification if we take into account the website structure. In the log file, we have accessed and referenced pages, so we can use heuristic methods for session identification. The H-ref method (Spiliopoulou, Mobasher, Berendt, & Nakagawa, 2003) beside the referrer field use also the website map. If we use the heuristic identification and find two consecutive records with the same IP field, which are not directly connected with hyperlink, we can say that we find access of two unique visitors sharing same IP address. The h-ref method combines this procedure with the time threshold.

Another not so commonly used method of session identification is the Reference Length (Cooley, Mobasher, & Srivastava, 1997; Cooley et al., 1999; Kapusta, Munk, & Drlik, 2012; Kapusta, Munk, & Drlík, 2012) method based on the web site searching model. That model and the model of user behaviour are fundamental to the correct aggregation of individual user's clicks to meaningful sessions that are sometimes referred to as transactions. We can organize individual web pages of the examined web site to three groups in term of the model: content pages, navigation (auxiliary) pages, multiple purpose pages.

The content pages can be defined as the web pages where we can find the required information. These pages are the reason of the user visit throughout his browsing of web space. Therefore, we can say that, in the case of association rules searching, content pages are the most important. Our objective is to discover useful rules among those pages. We can define the user session as the set of the auxiliary

references up to and including each content reference for a given user (Auxiliary – Content Transactions). Mining auxiliary-content transactions would essentially give the common traversal paths through the web site to a given content page. The process of web browsing can be summarized as visiting many navigation pages a finish on the content page. If the user land on the content page, the sessions will end. Forthcoming browsing or searching for another content page we define as the next session. The path to the content page through navigation pages as referenced as the transaction in the literature.

1.1 The aim of the experiment

As part of the research at the Department of Computer Science of Constantine the Philosopher University in Nitra, we conducted several experiments in which we evaluated the relevance of the various data preparation steps of web server log file for sequence analysis. We investigated which steps of data preparation for proper analysis of the website with anonymous access are important. The motivation for our current research were the following facts:

- Experiments we conducted in the past showed that the user session identification had a significant impact to results of WUM sequence analysis are significant impact.
- The most used method of user session identification is the STT method, as the literature review shows.
- Previous experiments we realized showed that the recommended STT values (10, 15 and 30 minutes) are not "ideal". The STT value can be set dynamically as presented in (Huynh & Miller, 2009; Štencl & Šťastný, 2009; J. Zhang & Ghorbani, 2004).
- There are few more experimental methods of user session identification, e.g. Reference Length method and H-ref method which both were not compared with "traditional" session identification methods yet.

We propose an experiment for user session identification with the aim to compare the STT method with the unambiguous cookie method. We also want to verify the suitability of other session identification methods as well as the appropriateness of two new STT calculation methods.

1.2 Related work

In this section, we describe already realized experiments where the experimental methods based on heuristics (web structure based heuristics, web page time spent heuristic) were used. To detect limits of the user session we can use time thresholds (Gayo-Avello, 2009). This method is popular as it is simple to implement, and it often used with various values of STT. Most common values of STT are 5 minutes (Downey, Dumais, & Horvitz, 2007), 10 minutes (Huang, Chien, & Oyang, 2003; Chien & Immorlica, 2005), 15 minutes (He & Harper, 2000) and 30 minutes (Radlinski & Joachims, 2005). Most commonly used values of STT were also the aim of research of data preprocessing from the virtual learning environment log file (Drlik & Skalka, 2011; Munk & Drlik, 2011). Author compared the results when the 15, 30 and 60 minutes STT where used in combination with/without path completion. Results showed the application of lower values of STT (15 and 30 minutes) to session identification has impact to lowering the amount of trivial and inexplicable rules.

We can also mention "non-standard" STT value of 25.5 minutes used in (Catledge & Pitkow, 1995). Authors use this value as the sum of the average time the user spent on the website and 150% of standard deviation of the time the user spent on the website. Based on these experiments, the STT value of 30 minutes becomes most common (Chakrabarti et al., 1999) even it not always most appropriate. Huynh and Miller realized experiments with dynamic STT based on the type of web service (Huynh & Miller, 2009). Zhang a Ghorbani (J. Zhang & Ghorbani, 2004) compared 10

minutes STT, 30 minutes STT and dynamic STT. The value of dynamic STT was based on the time a visitor spent on a page (the *Length* variable and its standard deviation). They experimented with various characteristics of the *Length*, and they excluded the zero spent time (web crawlers) and highest spent time. They compared overall 11 time oriented methods for the session identification.

The heuristic method H-ref used Seco and Cardoso (Seco & Cardoso, 2006). The record from the log file was marked as a new session in case of missing referral in previous sessions or if the time interval between to consecutive record is more than 60 minutes. We also use this value in proposed experiment.

2 Research methodology

When examining the impact of data pre-processing steps for the quality and quantity of the knowledge we proceed as follows:

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 matrix from the log file (access data) and sitemap (content data).
3. Data preparation.
4. Data analysis – discovering user behavioural patterns. For the extraction of sequence rules we use the Apriori algorithm implementation (Agrawal, Imielinski, & Swami, 1993).
5. Understanding data after the analysis.
   We create data file based on the output of the analysis and calculate basic characteristics of examined files: accesses count, number of identified sequences, number of frequent sequences, average length of identified sessions.
6. The comparison of knowledge we acquired from examined files (pre-processed with different level of data preparation)
   When evaluating the acquired knowledge, we focus not only on the quantity of extracted rules, but also on their quality. Quality sequence rules is assessed by two indicators (Berry & Linoff, 2004; Stankovicova, 2009): support and confidence.
   We evaluate the acquired knowledge in terms of quantity and quality of founded sequence rules in the mean of:
   a. Comparison of the portion of founded rules.
   b. Comparison of the support and confidence of founded rules.

2.1 Data description

The source of data for our experiment is web server log file of domestic significant commercial bank operating in Slovakia gathered from selected sub content according the to valid EU Directives and National Bank of Slovakia Decrees) for 4 years (2009 – 2012). The data set compromise a/ Pillar 3 Disclosure Requirements Quarterly and Semi-annually, b/ Pillar 3 related information, c/ Other information. Pillar 3 Disclosure Requirements consists of information on the Bank (e.g. organizational chart, number of employees, list of the bank activities (e.g. in respect to diversity management (Egerová, Jiřincová, Lančarič, & Savov, 2013)) etc.); financial information (financial statement information, information on asset quality, information on liquidity etc.); information on risk management (risk strategies, policies, credit risk management, on risk weighted assets according to risk measurement approaches etc.), information on securitisation etc. Pillar 3 related information are those which are contained in the bank annual reports, minutes from general assembly meetings, prospect of emitent, information about group and information for banks. Part Other information contains such information as the bank history, awards, mission-vision-values, anti money laundering, ethical codes, contacts, etc.

2.2 Data preparation

A web server log file was pre-processed in the standard way. When we excluded worthless data (requests for images, style sheets, etc.) and web crawlers' accesses, we have got 938497 records from the year 2010. Before we can apply session identification methods, we have to get the time a user spent on a particular website. If the time interval between two consequent records in the log file is more than 60 minutes, the consecutive record represents a new session. It is unlikely that some user read the page more than 60 minutes. The value of σ is set to 60.

User sessions were identified in following data preparation levels:

- (File REF-L1): session identification based on the *Reference Length* without path completion,
- (File REF-L2): session identification based on the *Reference Length* with path completion,
- (File STT-M1): session identification based on the *STT* size equal to average duration from all sessions without path completion,
- (File STT-M2): session identification based on the *STT* size equal to average duration from all sessions with path completion,
- (File STT-Q1): session identification based on the *STT* size equal to quartile from all sessions without path completion,
- (File STT-Q2): session identification based on the *STT* size equal to quartile from all sessions with path completion,
- (File H-REF1): session identification based on the H-ref method without path completion,
- (File H-REF2): session identification based on the H-ref method with path completion.

2.3 Algorithms used in the data preparation

In the next section, we describe results important for application of the session identification methods as well as individual algorithms for the session identification.

Reference Length

The Reference Length heuristic is based on the assumption that the time spent on site (*Length*) is related to whether the site is classified as content or navigation. Assuming that the variable *Length* has an exponential distribution, and we know the portion of navigation pages, we can use the quantile function to estimate the border time. In our case, the use of heuristics is desirable due to the fact that we are able to classify the page as navigation or content one. We are able to determine the portion of the navigation pages. Due the nature of the bank website we do not need to bother with mixed pages.

The bank portal consists from 364 pages including pdf and doc files. Document files are classified as content pages (273 pages). Other pages are classified as navigation ones (91 pages) in case that the page contains at least one hyperlink excluding the main menu. If the page in not classified as the navigation page we can say, it is content one.

Based on the frequency of navigation and content pages, the proportion of navigation pages is $p = 0.25$. We can use the Reference Length method only if the *Length* variable representing the time spent on a particular page is exponential distributed.

The zero hypothesis; the *Length* variable fit the supposed distribution; is rejected at the 1% significance level based on the results of Chi-Square test *(Chi-Square test* = 3502953.27155*, df* = 22 (adjusted)*, p* = 0.00000). The *Length* variable almost fit the supposed distribution, but there are some deviations from the exponential distribution.

If we made the assumption about the portion of the navigation pages in the log file, we would determine the threshold time *C*, which separates the content pages from the navigation ones. The value of the threshold time divide the navigation pages and content pages according the time a user spent on a particular page. The next session starts with the page, where $Length_k > 32\ sec.$ The first *k-1* pages are classified as navigation pages and the last page (*k*) is classified as the content one. The time spent on the last page is higher than the threshold time.

Estimation of session threshold time

If the *Length* variable represents the time a user spent on the page, we can estimate another value of the threshold time. We inspired with methods of the non-outlier range of the *Length* variable. We suppose that values higher than $QIII + 1.5Q$ represent the end of the session.

	Valid N	Mean	Median	Min	Max	Lower Q	Upper Q	Range	Quartile Range	$Q_{III} + 1.5Q$
Length	573186	111.0273	18	0	3600	8	55	3600	47	125.5

Table 1: Descriptive characteristics of the *Length* variable

Firstly we estimate the value of STT based on the average value of the *Length*, which is 112 seconds (Table 1). Files with identified sessions are called STT-M1 and STT-M2. The *Length* values higher than 126 seconds we consider as non-outliers and files with identified session are called STT-Q1 and STT-Q2.

H-ref algorithm

The H-ref method does not consider the time a user spent on a particular page, even there still exist the 60 minutes STT to maintain extreme cases. The algorithm uses the referrer field from the web server log file. We implemented this algorithm using the php+mysql in our experiment and got files with identified sessions H-REF1 and H-REF2.

3 Experiment Results

The most identified sessions were in REF-L files. In comparison to the lowest number of identified sequences in H-REF files, it is increased by 45 %. The lowest discrepancy in the number of identifies sessions was between the files STT-M and STT-Q. It is obvious that the minimum discrepancy in the number of sequences (6741) is caused by the lowest difference in STT (Session Timeout Threshold) between files STT-M and STT-Q. The number of frequented sessions has almost increased by 150 % between the lowest number of frequented sequences in file REF-L1 and the highest number of sequences in file H-REF2.

	Number of records	**Number of visits – identified sessions**	**Average length of identified sessions**	**Frequented sequences**
REF-L1	938497	531832	2	69
STT-M1	938497	423422	2	103
STT-Q1	938497	416681	2	105
H-REF1	938497	366074	3	159
REF-L2	967117	531832	2	72
STT-M2	983540	423422	2	106
STT-Q2	985082	416681	2	109
H-REF2	1005603	366074	3	173

Table 2: Number of visits and sequences in individual files

The number of records has naturally increased after paths completion. The H-REF method has recorded the highest increase i.e. the discrepancy between H-REF1 and H-REF2 is 67106 records, what represents the increase more than 7 %. The lowest increase was by using REF-L method where the discrepancy between REF-L1 and REF-L2 is 28620 records, what represents increase by 3 %.
Based on the initial results we articulated the following assumptions:

- We expect that the identification of the session using H-REF method will have a significant impact on quantity of extracted rules.
- Since H-REF method works with web map and principally tries to create the meaningful sessions, we expect that, in case of using H-REF method, the path completion will not have a significant impact on quality of extracted rules in terms of their basic characteristics of quality.
- We expect that the path completion using all examined methods will have a significant impact on the quantity of extracted rules.

3.1 Comparison of proportion of the rules found in examined files

The results of the analysis of the proportion of the fund rules are sequence rules. They were obtained from the frequented sequences fulfilling the minimum support (see the summary in Table 3). Similar as in the previous experiment, we set up the value on min $s = 0.005$. Frequent sequences were obtained from identified sessions, i.e. from the individual users' portal visits during the examined year 2010 .

	REF-L1	STT-M1	STT-Q1	H-REF1	REF-L2	STT-M2	STT-Q2	H-REF2
Sum	42.00	85.00	88.00	169.00	47.00	91.00	98.00	204.00
Percent 0's	79.90	59.33	57.89	19.14	77.51	56.46	53.11	2.39
Percent 1's	20.10	40.67	42.11	80.86	22.49	43.54	46.89	97.61
Cochran Q test					$Q = 679.4912$, $df = 7$, $p < 0.000000$			

Table 3: Summary of the extracted sequence rules in individual files

The most rules were extracted from the files H-REF in case of path completion as well as without path completion. In comparison to the lowest number of rules in file REF-L, it is an increase by 162 rules, which represents 385 % of extracted rules. Through that, the more rules were extracted in examined files after path completion (REF-L2, STT-M2, STT-Q2, H-REF2) than without path completion (REF-L1, STT-M1, STT-Q1, H-REF1) this increase was not significant. It is increasing of number of rules from 7 % (STT-M1 vs. STT-M2) to 20 % (H-REF1 vs. H-REF2). The highest increase of the rules after path completion was surprising for us. The principle of the method consists of the meaningful sessions searching based on the web map. For this reason, we did not expect the large increase of the number of extracted rules after path completion. Therefore, it will be interesting to compare it in terms of quality of extracted rules.

Results of sequence analysis show the high concordance of proportion of extracted rules between files STT-M and STT-Q after path completion as well as without path completion. This concordance is natural, since it is an application of STT with similar sizes.

Based on the results of Q test (Table 3), the zero hypothesis, which reasons that the occurrence of the extracted rules does not depend on individual levels of data preparation for WUM, is rejected at the 1 % significance level. Kendall's coefficient of concordance represents the degree of concordance in the number of extracted rules in individual files. The value of the coefficient is 0.46, where 1 means a perfect concordance and 0 represent discordance. Low value of the coefficient confirms Q test result.

Occurrence	Mean	1	2	3	4
File(REF-L1)	0.200957		****		
File(REF-L2)	0.224880		****		
File(STT-M1)	0.406699	****			
File(STT-Q1)	0.421053	****			
File(STT-M2)	0.435407	****			
File(STT-Q2)	0.468900	****			
File(H-REF1)	0.808612			****	
File(H-REF2)	0.976077				****
Kendall Coeff. of Concordance					**0.46445**

Table 4: Homogenous groups in examined files

From the multiple comparison (Tukey HSD test) four homogenous groups (Table 4) were identified. The largest group consists of files STT-M1, STT-M2, STT-Q1 and STT-Q2. Path completion has no significant impact on the quantity of extracted sequence rules in terms of identification of session using STT methods. Statistically significant differences were proved between STT methods and others.

Next identified group, in terms of the average occurrence of extracted rules, consists of files where the sessions were identified using Reference Length with/without application of path completion. Statistically significant differences at the 5 % significance level were proved between group of files with application of H-REF method and others and also between the files with application of H-REF method with path completion (H-REF2) and without path completion (H-REF1).

Path completion has a significant impact on the quantity of extracted rules in case of the use of H-REF method. Likewise, the application of H-REF method has a significant impact on the quantity of extracted rules.

3.2 Comparison of quality of the rules found in examined files

In the previous section, we identified homogeneous groups in terms of average occurrence of extracted rules. i.e. in terms of quantity of extracted rules. More comprehensive view is obtained from the comparison of quality of sequence rules assessed by indicators- support and confidence. There are differences in results of sequence rule analysis among individual files in terms of values of support and confidence of extracted rules.

Support	Mean	1	2	3
File(REF-L1)	1.511783		****	
File(REF-L2)	1.563045		****	
File(STT-M1)	2.515048	****		
File(STT-M2)	2.572087	****		
File(STT-Q1)	2.592141	****		
File(STT-Q2)	2.648962	****		
File(H-REF1)	3.260186			****
File(H-REF2)	3.314136			****
Kendall Coeff. of Concordance				**0.87589**

Table 5: Homogenous groups for characteristics – support of extracted rules

Kendall's coefficient of concordance represents the degree of concordance in support of extracted rules among the examined files (Table 5). The value of the coefficient (Table 5) is 0.88, where 1 means a perfect concordance and 0 represent discordance.

From the multiple comparison (Tukey HSD test), three homogenous groups (Table 6) were identified in both cases. Statistically significant differences at the 5 % significance level in support and confidence of extracted rules were proved among files with identification of sessions using STT methods (STT-M1, STT-M2, STT-Q1, STT-Q2) and others and also among files with identification of sessions using H-REF methods (H-REF1, H-REF2) and others. The basic difference among quantity and quality of extracted rules is that, in case of quality comparison, statistically significant difference was not proved in method of session identification using H-REF method without path completion (H-REF1) and using H-REF method with path completion (H-REF2).

Confidence	Mean	1	2	3
File(REF-L1)	10.17628		****	
File(REF-L2)	10.32179		****	
File(STT-M1)	13.17686	****		
File(STT-M2)	13.31388	****		
File(STT-Q1)	13.37286	****		
File(STT-Q2)	13.51686	****		
File(H-REF1)	14.92257			****
File(H-REF2)	15.08428			****
Kendall Coeff. of Concordance				**0.73645**

Table 6: Homogenous groups for characteristics – confidence of extracted rules

4 Conclusion

The session identification based on the Reference Length is conditioned with the exponential distribution of the *Length*, which, however, does not have it in our case. The need of the exponential distribution of the *Length* seems to be the biggest problem of this method. There is a probability that even small deviations from the exponential distribution affect the correct calculation of the border time. The experiment was carried out on the log file of the commercial bank in the year 2010. Before running the experiment, we examined the distribution of the *Length* at the university portal as well as the newer log files of the bank portal. In both cases, the exponential distribution of the *Length* was not confirmed.

The second problem of the Reference Length method appears to be the portion of navigation pages. It is obvious that, with increasing age of internet portal, the number of content pages increases more than navigation ones. If we take a look on news portals, the navigation pages with links to stories categories grow slower than pages with stories itself. The border time C depends on the portion of navigation and contents pages. From this perspective, the portion significantly limits the use of this method.

The results of the H-ref method were disappointing. The methods itself try to create meaningful sessions. We improved the method by employing the path completion. Even after this enhancement, there are still statistically significant differences in the quantity of extracted rules when using the path completion or not. If we take a look on the quality of rules, there are no statistical differences. The method can be used for portals that are characteristic with flood visitors, e.g. e-auctions, e-tickets.

The best results were achieved by employing user session identification methods based on the threshold time (STT). We think that if we want to achieve the best results regarding the quality of pre-

processed data, we have to use the method of STT to the log file. The value of STT is calculated from formula:

$$Date_{i+1} - Date_i > Q_{III} + 1.5Q, \quad (1)$$

where $Date_i$ stand for the last record with the web site access time in investigated session,
$Date_{i+1}$ stand for the first record with the access time in new session,
Q_{III} stand for the last value of upper quartile of the *Length* variable
and Q stand for the quartile range of the *Length* variable.

The experiment proved that the session identification based on STT calculated according the formula (1) is the most appropriate for gathering the sequence of pages visited by a particular user.

Getting the value of the *Length* variable as well as the calculation of quartile ranges can be in most cases problematic in technological, economic a time view. The experiment proved that comparable results are obtained by identifying the sessions using STT with a value equal to the average of the *Length*, i.e. the average time that the user spent on the examined pages. This value is usually available in standard systems for the web analysis.

Acknowledgements 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

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.*, *22*(2), 207–216

Bayir, M. A., Toroslu, I. H., & Cosar, A. (2006). A New Approach for Reactive Web Usage Data Processing. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on* (p. 44)

Berendt, B., & Spiliopoulou, M. (2000). Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, *9*, 56

Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. *portal.acm.org* (p. 545)

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems*, *27*, 1065–1073

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Grouping Web page references into transactions for mining World Wide Web browsing patterns. *Proceedings 1997 IEEE Knowledge and Data Engineering Exchange Workshop*

Cooley, R., Mobasher, B., Srivastava, J., & others. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, *1*, 5–32

Downey, D., Dumais, S., & Horvitz, E. (2007). Models of Searching and Browsing: Languages, Studies, and Application. *IJCAI*, 2740–2747

Drlik, M., & Skalka, J. (2011). Virtual faculty development using top-down implementation strategy and adapted ees model. *Procedia - Social and Behavioral Sciences*, *28*, 616–621

Egerová, D., Jiřincová, M., Lančarič, D., & Savov, R. (2013). Applying the concept of diversity management in organisations in the Czech Republic and the Slovak Republic - a research survey. *Technological and Economic Development of Economy*, *19*(2), 350–366

Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, *179*, 1822–1843

He, D., & Harper, D. J. (2000). Detecting session boundaries from Web user logs. In *Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research* (pp. 57–66).

Huang, C.-K., Chien, L.-F., & Oyang, Y.-J. (2003). Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*, *54*(7), 638–649

Huynh, T., & Miller, J. (2009). Empirical observations on the session timeout threshold. *Information Processing and Management*, *45*, 513–528

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J. (1999). Mining the Web's link structure. *Computer*, *32*

Chen, Z., Fu, A. W.-C., & Tong, F. C.-H. (2003). Optimal algorithms for finding user access sessions from very large web logs. *World Wide Web*, *6*(4), 259–279

Chien, S., & Immorlica, N. (2005). Semantic Similarity Between Search Engine Queries Using Temporal Correlation. *Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan*, 2–11

Kapusta, J., Munk, M., & Drlik, M. (2012). User Session Identification Using Reference Length. In *DIVAI 2012: 9TH INTERNATIONAL SCIENTIFIC CONFERENCE ON DISTANCE LEARNING IN APPLIED INFORMATICS: CONFERENCE PROCEEDINGS* (pp. 175–184).

Kapusta, J., Munk, M., & Drlík, M. (2012). Cut-off time calculation for user session identification by reference length. In *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*

Klocoková, D. (2011). Integration of heuristics elements in the web-based learning environment: Experimental evaluation and usage analysis. *Procedia - Social and Behavioral Sciences*, *15*, 1010–1014

Koprda, Š., Balogh, Z., & Turčáni, M. (2011). Fuzzy control rules base design. In *2011 5th International Conference on Application of Information and Communication Technologies, AICT 2011*

Li, Y. L. Y., Feng, B. F. B., & Mao, Q. M. Q. (2008). Research on Path Completion Technique in Web Usage Mining. *2008 International Symposium on Computer Science and Computational Technology*, *1*

Munk, M., & Drlik, M. (2011). Influence of Different Session Timeouts Thresholds on Results of Sequence Rule Analysis in Educational Data Mining. In *Digital Information and Communication Technology and Its Applications, Pt I* (Vol. 166, pp. 60–74)

Munk, M., & Drlík, M. (2014). Analysis of stakeholders' behaviour depending on time in virtual learning environment. *Applied Mathematics and Information Sciences*, *8*(2), 773–785

Nasraoui, O., & Saka, E. (2007). Web usage mining in noisy and ambiguous environments: Exploring the role of concept hierarchies, compression, and robust user profiles. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4737 LNAI, pp. 82–101)

Nina, S. P., Rahman, M., Bhuiyan, K. I., & Ahmed, K. (2009). Pattern Discovery of Web Usage Mining. In *2009 International Conference on Computer Technology and Development* (Vol. 1)

Olej, V., & Filipova, J. (2012). Short time series of website visits prediction by RBF neural networks and support vector machine regression. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7267 LNAI, pp. 135–142)

Olej, V., & Filipová, J. (2011). Modelling of web domain visits by radial basis function neural networks and support vector machine regression. In *IFIP Advances in Information and Communication Technology* (Vol. 364 AICT, pp. 229–239)

Radlinski, F., & Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. In G. R.L., B. R., B. K., & V. J. (Eds.), *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 239–248)

Seco, N., & Cardoso, N. (2006). *Detecting User Sessions in the Tumba! Query Log*.

Song, J. S. J., Luo, T. L. T., Chen, S. C. S., & Gao, F. G. F. (2007). The Data Preprocessing of Behavior Pattern Discovering in Collaboration Environment. *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS Journal on Computing*, *15*, 171–190

Stankovicova, I. (2009). Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner. In *Informační a datová bezpečnost ve vazbě na strategické rozhodování ve znalostní společnosti* (pp. 1–7).

Štencl, M., & Štastný, J. (2009). Advanced approach to numerical forecasting using artificial neural networks. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *57*, 297–304

Tan, S.-H., Chen, M., & Yang, G.-H. (2010). User behavior mining on large scale web log data. *The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding*, 60–63

Turcinek, P., Stastny, J., & Motycka, A. (2012). Usage of Data Mining Techniques on Marketing Research Data. In *Proceedings of the 11th WSEAS International Conference on Applied Computer and Applied Computational Science* (pp. 159–164)

Zhang, C., & Zhuang, L. (2008). New Path Filling Method on Data Preprocessing in Web Mining. *Computer and Information Science*, *1*(3), 112–115

Zhang, J., & Ghorbani, A. A. (2004). The reconstruction of user sessions from a server log using improved time-oriented heuristics. *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004*