ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LXI 108 Number 4, 2013
http://dx.doi.org/10.11118/actaun201361040973

DATA PRE-PROCESSING FOR WEB LOG MINING: CASE STUDY OF COMMERCIAL BANK WEBSITE USAGE ANALYSIS

Jozef Kapusta, Anna Pilková, Michal Munk, Peter Švec

Received: April 11, 2013

Abstract

KAPUSTA JOZEF, PILKOVÁ ANNA, MUNK MICHAL, ŠVEC PETER: *Data pre-processing for web log mining: Case study of commercial bank website usage analysis.* Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 2013, LXI, No. 4, pp. 973–979

We use data cleaning, integration, reduction and data conversion methods in the pre-processing level of data analysis. Data processing techniques improve the overall quality of the patterns mined. The paper describes using of standard pre-processing methods for preparing data of the commercial bank website in the form of the log file obtained from the web server. Data cleaning, as the simplest step of data pre-processing, is non–trivial as the analysed content is highly specific. We had to deal with the problem of frequent changes of the content and even frequent changes of the structure. Regular changes in the structure make use of the sitemap impossible. We presented approaches how to deal with this problem. We were able to create the sitemap dynamically just based on the content of the log file. In this case study, we also examined just the one part of the website over the standard analysis of an entire website, as we did not have access to all log files for the security reason. As the result, the traditional practices had to be adapted for this special case. Analysing just the small fraction of the website resulted in the short session time of regular visitors. We were not able to use recommended methods to determine the optimal value of session time. Therefore, we proposed new methods based on outliers identification for raising the accuracy of the session length in this paper.

association rules, web log mining, business intelligence, financial regulation, market discipline, data preprocessing methodology

During the recent financial crisis, the Market discipline (MD) has been recognized as an important supplement of regulators´ supervisory efforts in banking. We analyse the market participants' interest in mandatory disclosure of financial and risk information by a commercial bank by means of advanced methods of web log mining. We ascertain whether the purposes of Basel 2, Pillar 3 have been fulfilled. According to the information released, stakeholders can consider banks' risk exposure, management and capital adequacy.

Main source of information for web log mining is the web server log file. We use data from the public commercial bank web site which cover a period from the fourth quarter of the 2008 to the fourth quarter of the 2010. The period was selected by design as it covers website accesses before and during the recent financial crisis. A common web server keeps accesses of user in the log file and logs basic information about user's computer (e.g. IP address, date and time referring page, browser version). Logs provide basic data as they record page accesses, not interaction with the page, and they cannot make relevant distinctions such as distinguishing between the time a user spends reading and the time they spend away from the screen (Thomas, 2012). Data of outstanding quality requires rigorous data gathering as the data preprocessing. Data preparation is probably the longest and most time-consuming phase in the process of the web usage mining. The reason is the incompleteness of accessible data as well as irrelevant information present in the collected data.

In the following section, we describe methodologies how to reconstruct the activity of every user, how to detach his activities from activities of other users while preserving his anonymity. This is demanding procedure from the perspective of theory, time and the technical realisation. Different aspects of data preparation and modelling can be found, for example, in work by Chitraa and Davamani (2010), Liu, Mobasher and Nasraoui (2011), Bing (2006), Koprda (2012), Houšková (2011), Popelka and Šťastný (2009) or Fejfar and Šťastný (2011). We also try to improve methods of session time threshold as this is key variable in session identification.

RESEARCH METHODOLOGY

Basic data preprocessing

Experiments that analyse user's behaviour usually take smaller set of data, e.g. two weeks or one month (Liu and Kešelj, 2007; Stevanovic, An and Vlajic, 2012). Another approach is to select few reference weeks (Xing and Shen, 2004). We analyse a quite extensive period (27 months), so we have to change common used methods (Munk, Kapusta, and Švec, 2010; Munk, Kapusta, and Švec, 2009; Munk and Drlík, 2011a) and the internal applications for log file preprocessing. We have to rewrite these desktop applications to the server ones as it can run in batch mode.

First step in the log file preprocessing is the removal of unnecessary data. These data represents access to graphic files or style sheets (Ortega and Aguillo, 2010) so the accesses from web robots. Web robots may be defined as autonomous systems that send requests to web servers across the Internet. A canonical example of a web robot is a search engine indexer. A less common example is an RSS feed crawler or a robot designed to collect web sites for an Internet archive (Doran and Gokhale, 2011). The difference between the robot and the human can be determined based on basic and common features like click rate, HTML-to-Image ratio, percentage of file requests, percentage of 4xx error responses, percentage of HEAD requests, or access to robot.txt file. We can use new methods like standard deviation of requested page's depth or percentage of consecutive sequential HTTP requests (Stevanovic, An and Vlajic, 2011). There are many robots or crawlers that cannot be identified based on general crawler attributes. In this case, we can use the method navigational patters analysis (Tan and Kumar, 2002). After the cleaning of the log file and removing web crawlers' accesses and unnecessary data, the log file with just the 10% of the original file length, about 4 million of records.

Log files also collect information about IP address. In cooperation with the bank, we marked IP addresses that are used in the bank local network. A staff is more familiar with the organisation and has better knowledge of the website structure (Thomas and Paris, 2010). We removed these accesses because the majority of accesses are from content creators, web administrators and managers responsible for information disclosure.

Additional attributes to log file

The preliminary step for the data analysis is to mark what content categories are represented in the sitemap. The sitemap itself is well generated with the content management system of the bank website. In the case of the unavailability of this feature, we can use many free tools to create the sitemap. In the second step, bank expert is needed to mark sets of similar content from every page in the sitemap. Content sets represent the content categories in the analysis. Our analysed sitemap contained 68 parts of the content from which expert determined 23 categories, e.g. History, Financial data, Documents or Service charges. Analysed website use the search engine friendly URLs, so each page name is part of a hypertext link found in the log file.

The next step of an expert is to assign the affiliation of each category to Pillar 3. The group of *Pillar 3 Disclosure requirements* contained two categories and the group *Pillar 3 Related* contained 7 categories. Unmarked categories represent the *Other* group. Each category has also the financial / non-financial attribute. The mapping of URL to category can be seen in the Tab. I.

There are many tools that are able to analyse the web server log file, but we cannot used them for several reasons. We analyse the period of 3 years and the bank changed the syntax of the log file several times. The well-known structure of common log file cannot be used in the log file analysis. We have to create a tool for the log file analysis, which will be able to identify the change of the log file syntax, take appropriate action and properly assign defined observed variables. The tool also assigns the category, Basel category and the type of financial information. Based on the date information we also calculate the financial quarter (*QI, QII, QIII, QIV*) of the year to simplify the process of statistical analysis.

I: *Mapping URLs to categories*

URL (from the log file)	Category	Basel group	Information
/about/branch.html	Contacts	Other	Nonfinancial
/about/The-economic-results	Financial reports	Pillar 3 Related	Financial
/about/vysledky_banky/info_kvartal	Quarterly information	Pillar 3 Disclosure Requirements	Financial
/about/contacts/	Contacts	Other	Nonfinancial

There are also URLs that cannot be assigned to any from defined categories. In this case, we have to check if the referrer to this URL has the category assigned. The situation of missing URL assignment to category can be found in the request of PDF files. The bank website puts documents for download in the separate content as is analysed, but information tight to Pillar 3 are often published as PDF documents. The algorithm itself can be seen in the Fig. 1.

```
Foreach record from the logfile {
        Parse the record;
        Get category from URL;
        If not exist category {
        Get category from REFERRER;
        }
        If exist category {
                Assign basel_group based on category;
                Assign information based on category;
                Calculate Quartal from Timestamp;
                Get language from URL;
        }
        Assign category value -1
}
```

1: *Log file analysis algorithm*

Session identification

Session refers to a series of links made by one web-user. At present, there are some arithmetic for session identification – Hvist, where user's access time to the whole website will be given an upper limit, usually 30 min, Hpage where users's access time to one page will be given an upper limit, usually 10 min, and Href which is classified according to user's access history and reference pages (Fang and Huang, 2010).

For the session identification, we use Session Timeout Threshold (STT) (Munk and Drlik, 2011b). Other common methods are identification based on Access Time Threshold and session reconstruction (Fang and Huang, 2010). The aim of session identification is to divide accesses of every user into separate sets. The STT method divides the session into smaller ones if it finds accesses to the web page from the same source with higher interleave than the set period – time window. The appropriate size of the time window evolves the quality and quantity of found behaviour patterns.

Reference Length Method

The quantity and the quality of found behaviour patterns depend on the length of the STT. In our past studies, we used the average time visit as the STT. Incorrect estimation of STT can reduce (in case of low value of STT) or raise (in case of high value of STT) of found users behaviour patterns.

The more precise method for estimation of STT is the Reference Length method. The model of web site searching and model of user behaviour is fundamental to correct aggregation of the individual user's clicks to meaningful sessions or transactions. We can organize individual web pages of the examined web site to three groups in term of model – content pages, navigation (auxiliary) pages and multiple purpose pages. The content pages are web pages where the user can find required information. These pages are the reason of visit of the individual user throughout his browsing of web space. When we are searching association rules (Klocokova, 2011), content pages are most relevant. Our objective is to discover useful rules among those content pages.

A transaction can be defined as all of the auxiliary references up to and including each content reference for a given user (Auxiliary – Content Transactions). The mining of auxiliary-content transactions would essentially give the common traversal paths through the web site to a given content page. The reference length method is based on the assumption that the amount of time a user spent on a web page correlates to whether the web page should be classified as an auxiliary or content page for that user. Qualitative analysis of several other server logs reveals the shape of the histogram has a large exponential component (Cooley, Mobasher and Srivastava, 1999).

If we defined the assumption about the portion of navigation pages in surveyed log file, we can define the cut-off time *C* that separates the content pages and other types of pages.

When the cut-off time *C* is known the session can be created in such manner that we compare the time of a particular web page visit with the cut-off 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 spend there more time than *C*). We can claim the content page is the last page of session. The subsequent page is the first page of a new session.

The calculation of cut-off time *C* is the most important if we want to use the reference length transaction method for user session identification. The verification of the exponential distribution of variable *RLength* obtained from the log file is also coessential.

We assume that the variance of the times spent on the auxiliary pages is small because the visitor only goes through them with the objective to find required information on the content page. Therefore, the auxiliary references make up the lower end of the curve (Fig. 2). The variance of the times spent on the content pages is wide, and we assume that they make up the upper tail that extends out to the longest reference. If the assumption about the proportion of navigation pages in the log file exists, we can calculate the cut-off time *C* that divides web pages into navigation pages and content pages. We do not reject the null hypothesis

which claims that the variable *RLength* has assumed distribution.

The variable RLength has an exponential distribution.

$$f(RLength) = \lambda e^{-\lambda.RLength}, \quad (1)$$

$$F(RLength) = 1 - e^{-\lambda.RLength}, \quad (2)$$

where
$RLength \geq 0$.

If p is the relative frequency of navigation pages, we can apply the fractile function (inverse distribution function) to estimate cut-off time C.

$$F^{-1}(p,\lambda) = C = \frac{-\ln(1-p)}{\lambda}, \quad (3)$$

where $0 \leq p < 1$.

Maximum likelihood estimation of parameter λ (mean intensity of events) is

$$\hat{\lambda} = \frac{1}{\overline{RLength}}, \quad (4)$$

where
$\overline{RLength}$ is observed mean length of visits. (Inverted value of mean time spent on the web pages).

RESULTS

The Reference Length Method is based on the assumption that, time the user spent on the website (*Length*) depends on the classification (content or navigational) of the page. If we assume that the *Length* variable is exponentially distributed, and the ratio of navigational pages is known, we can use the quantile function to calculate the cut-off time. The calculated cut-off time can used to identify users' session time. The variance of time spent on the navigational pages is low as we see on the left part of the graph at Fig. 2.

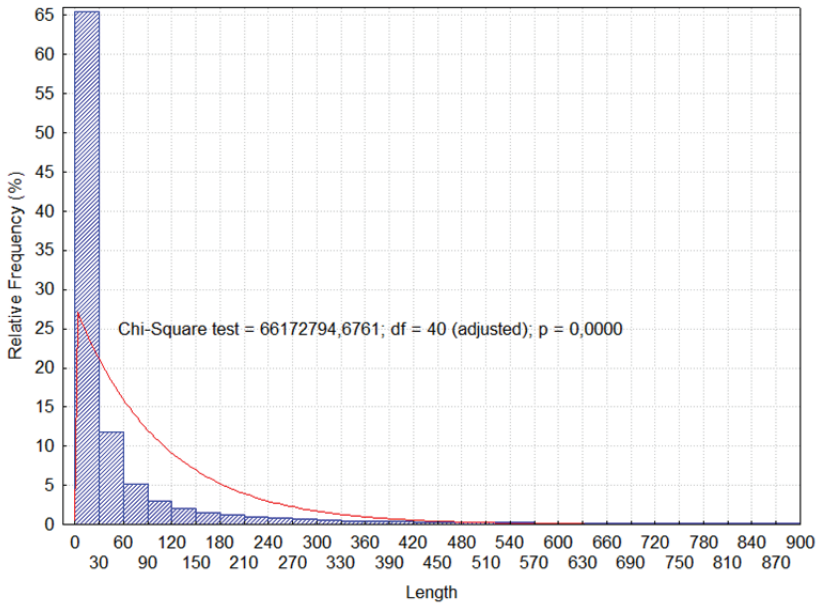
Based on the results of Chi-Square test (Fig. 2) the zero hypothesis is rejected at the 1% significance level. The *Length* variable does not have the exponential distribution so we cannot use the Reference Length method. We have to use the STT method where we can use the *Length* variable as the visit time. We can also calculate the Median, quartile range and non-outlier range of Length variable (Fig. 3).

Non-outlier range is created between the last value of QIII + 1,5Q and the last value of Length above QI–1,5Q. The values outside this interval we consider as outliers.

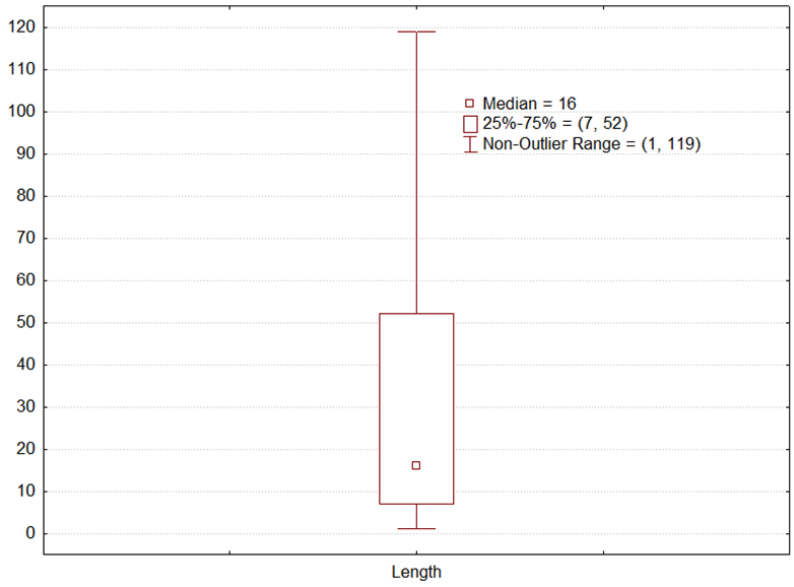
We are focusing on the upper limit of non-outlier range. We were inspired in cognitive science by applying these statistics (Stranovska *et al.*, 2012; Munkova, Stranovska and Durackova, 2012). Values those are higher than 119 are considered as outliers (Tab. II). The number 119 is the size of the time window. When the time between accesses to website from the same source is longer, the new session begins. We can also see that there are 85% values of the Length which are smaller than 120 seconds (Fig. 4).

DISCUSSION

In this paper, we pay attention to several steps in data pre-processing for the web log mining of the bank website before and during the last financial crisis. As this log file represents access from 3 years, we have to deal with new problems in the log file



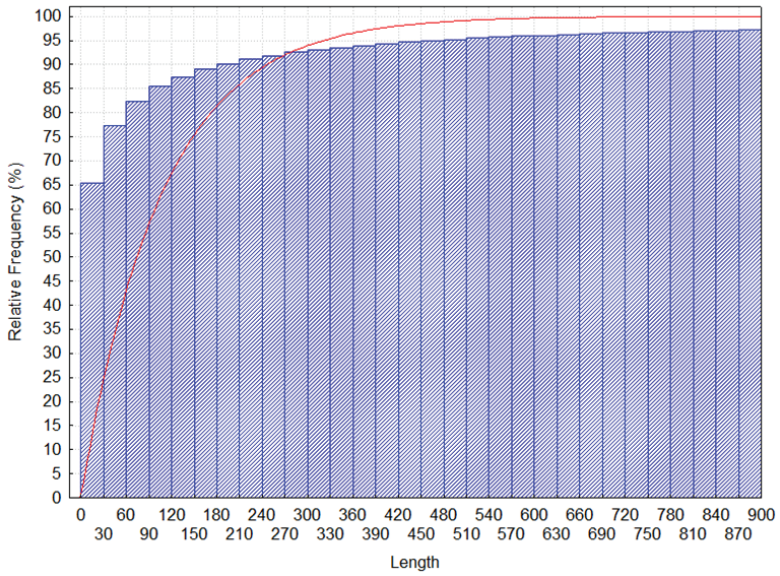
2: *Distribution of Length variable*



3: *Box plot: median/quartile range/non-outlier range of the Length variable*

II: *Descriptive characteristics of variable Length*

	Mean	Median	Minimum	Maximum	QI	QIII	QIII+1,5Q	Q
Length	107,17	16,00	1,00	3600,00	7,00	52,00	119,50	45,00



4: *Cumulative frequencies of Length variable*

analysis, mainly frequent changes of the log file syntax and the time demanding of the analysis itself. We have to add new variables based on the information found in the log file as these variables are needed for the analysis of the Pillar 3 information disclousure. Bank expert had to identify content categories for the analysis.

We also identify users sessions based on two different methods – the session time threshold and the reference length method. We try to estimate the STT with the use of Reference Length and take a closer look to the time variable which represent the duration of user visit. The results of the analysis show that the Length variable, which is essential for the Reference Length method, does fit to an exponential distribution. We have to use own method for the STT, and we used the non-outlier range. Non-outlier range is created between the last value of QIII + 1,5Q and the last value of Length above QI–1,5Q. The values outside this interval we consider as outliers. We are focusing on the upper limit of non-outlier range.

SUMMARY

We analyse domestic and foreign market participants' interests in mandatory Basel 2, Pillar 3 information disclosure of a commercial bank during the recent financial crisis. We try to ascertain whether the financial regulation and market discipline have been fulfilled. We model bank visitor behaviour based on the analysis of web log. We use data cleaning, integration, reduction and data conversion methods in the pre-processing level of data analysis. We propose new methods based on outliers identification for raising the accuracy of the session length. The analysis can help better understand the rate of use of the mandatory and optional Pillar 3 information disclosure. The results show that there is in general a small interest of stakeholders in mandating the commercial bank disclosure of financial information.

Acknowledgement

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

BING, L., 2006: *Web Data Mining. Exploring Hyperlinks, Contents and Usage Data*. Berlin: Springer-Verlag, 532 p. ISBN 978-3-540-37881-5.

COOLEY, R., MOBASHER, B. and SRIVASTAVA, J., 1999: Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*, 1, 1: 5–32. ISSN 0219-1377.

DORAN, D. and GOKHALE, S. S., 2011: Web robot detection techniques: overview and limitations. *Data Mining and Knowledge Discovery*, 22, 1–2: 183–210. ISSN 1384-5810.

FANG, Y. and HUANG, Z., 2010: An Improved Algorithm for Session Identification on Web Log. In: WANG, F. et al. (ed.). *Web Information Systems and Mining – LNCS*, 63, 18: 53–60. ISSN 0302-9743.

FEJFAR, J. and ŠŤASTNÝ, J., 2011: Time Series Clustering in Large Data Sets. *Acta Univ. Agric. et Silvic. Mendel. Brunen.*, 64, 2: 75–80. ISSN 1211-8516.

HOUŠKOVA BERANKOVÁ, M and HOUŠKA, M., 2011: Data, information and knowledge in agricultural decision-making. *Agris On-line Papers in Economics and Informatics*, 3, 2: 74–82. ISSN 1804-1930.

CHITRAA, V. and DAVAMANI, A., 2010: A Survey on Preprocessing Methods for Web Usage Data. *International Journal of Computer Science and Information Security*, 7, 3: 78– 83. ISSN 1947-5500.

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. ISSN 1877-0428.

KOPRDA, Š., TURČÁNI, M. and BALOGH, Z., 2012: Modelling, simulation and monitoring the use of LabVIEW. In: *6th International Conference on Application of Information and Communication Technologies, AICT 2012 – Proceedings*. Tbilisi: IEEE, 450–454. ISBN 978-1-4673-1740-5.

LIU, H. and KEŠELJ, V., 2007: Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data & Knowledge Engineering*, 61, 2: 304–330. ISSN 0169-023X.

LIU, B., MOBASHER, B. and NASRAOUI, O., 2011: Web Usage Mining. In: *Web Data Mining*. Berlin: Springer, 527–603. ISBN 978-3-642-19459-7.

MUNK, M. and DRLÍK, M., 2011a: Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System. *Procedia Computer Science*, 3, 4: 1640–1649. ISSN 1877-0509.

MUNK, M. and DRLÍK, M., 2011b: Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining. *Communications in Computer and Information Science*, 166: 60–74. ISSN 1865-0929.

MUNK, M., KAPUSTA, J. and ŠVEC, P., 2010: Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor. *Procedia Computer Science*, 1, 1: 2273–2280. ISSN 1877-0509.

MUNK, M., KAPUSTA, J. and ŠVEC, P., 2009: Data preprocessing dependency for web usage mining based on sequence rule analysis. In: *IADIS Multi Conference on Computer Science and Information Systems*. Algarve: MCCSIS, 179–181. ISBN 978-972-8924-88-1.

MUNKOVA, D., STRANOVSKA, E. and DURACKOVA, B., 2012: Impact of Cognitive-Individual Variables on Process of Foreign Language Learning. *Procedia Social and Behavioral Sciences*, 46: 5430–5434. ISSN 1877-0428.

ORTEGA, J. L. and AGUILLO, I., 2012: Differences between web sessions according to the origin of their visits. *Journal of Informetrics*, 4, 3: 331–337. ISSN 1751-1577.

POPELKA, O. and ŠŤASTNÝ, J., 2009: WWW Portal Usage Analysis Using Genetic Algorithms. *Acta Univ. Agric. et Silvic. Mendel. Brunen.*, 62, 6: 201–208. ISSN 1211-8516.

STEPHANOU, C., 2012: Rethinking market discipline in banking: lessons from the financial crisis. *Policy research working paper*. ISSN 1813-9450.

STEVANOVIC, D., AN, A. and VLAJIC, N., 2011: Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers. In: KRYSZKIEWICZ, M. et al. (ed.). *Foundations of Intelligent Systems – LNCS,* 6804: 483–489. ISSN 0302-9743.

STEVANOVIC, D., AN, A. and VLAJIC, V., 2012: Feature evaluation for web crawler detection with data mining techniques. *Expert Systems with Applications*, 39, 10: 8707–8717. ISSN 0957-4174.

STRANOVSKA, E., FRATEROVA, Z., MUNKOVA, D. and MUEGLOVA, D., 2012: Politeness factors in requests formulated in the 'category width' cognitive style. *Studia psychologica*, 54, 2: 111–124. ISSN 0039-3320.

TAN, P.N. and KUMAR, V., 2002: Discovery of Web Robot Sessions Based on their Navigational Patterns. *Data Mining and Knowledge Discovery*, 6, 1: 9–35. ISSN 1384-5810.

THOMAS, P., 2012: Explaining difficulty navigating a website using page view data, In: *Proceedings of the Seventeenth Australasian Document Computing Symposium 2012*, ACM: Dunedin, 31–38, ISBN 978-1-4503-1411-4.

THOMAS, P. and PARIS, C., 2010: Interaction differences in web search and browse logs. In: *Proceedings of the 15th Australasian Document Computing Symposium2010,* Melbourne: RMIT University, 52–60. ISBN 978-1-921426-80-3.

XING, D. nad SHEN, J., 2004: Efficient data mining for web navigation patterns. *Information and Software Technology*, 46, 1: 55–63. ISSN 0950-5849.

Address

PaedDr. Jozef Kapusta, PhD., Department of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia, doc. Ing. Anna Pilková, CSc., MBA, Department of Strategy and Entrepreneurship, Commenius Univeristy in Bratislava, Šafárikovo nám. 6, 818 06 Bratislava, Slovakia, doc. RNDr. Michal Munk, PhD., Department of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia, PaedDr. Peter Švec, Ph.D., Department of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia, e-mail: jkapusta@ukf.sk,anna.pilkova@fm.uniba.sk, mmunk@ukf.sk, psvec@ukf.sk