Experimental Verification of the Dependence Between the Expected and Observed Visit Rate of Web Pages

Jozef Kapusta(✉), Michal Munk, and Martin Drlik

Constantine the Philosopher University in Nitra,
Tr. A. Hlinku 1, 949 74 Nitra, Slovakia
{jkapusta,mmunk,mdrlik}@ukf.sk

Abstract. This paper is focused on a utilization of the web usage mining and web structure mining methods. We tried to answer the question if the expected visit rate of individual web pages correlates with the observed visit rate of the same web pages. We used web server log files as a data source. We applied several log file pre-processing methods to identify the user sessions on different levels of granularity. We found out that the quality of acquired knowledge about the users' behaviour depends on the method of the session identification. We have experimentally proved a higher dependence between the observed and expected visit rates of the examined web pages in well-prepared files with identified user sessions. We found out statistically significant differences between PageRank and a real visit rate in the files with application of more advanced methods of session identification.

Keywords: Web usage mining · Web structure mining · PageRank · Support · Observed visit rate · Expected visit rate

1 Introduction

The aim of the website designers or creators is to provide information to users in a clear and understandable form. Information displayed on individual web pages is interconnected by hypertext references. The website creator can affect visitors' behaviour by defining of references between web pages. He indicates the importance of information displayed on web pages through these references. It is probably true more references head to more important web pages. These are directly accessible from the home page or are referred from other important web pages.

Web pages are mostly understood as an information resource for users. They can also provide information in an opposite direction. The website providers can collect information about their users or about users' behaviours, needs or interests.

The knowledge discovery from the web page structure is known as a web structure mining (WSM) [1]. From WSM point of view, we focused on an analysis of quality and importance of web pages based on the references (links) among web pages. Determination of the web page importance is based on the idea that the degree to which we can rely on the web page quality is transferred by the references to web pages. If the

D.-S. Huang and K. Han (Eds.): ICIC 2015, Part III, LNAI 9227, pp. 637–648, 2015.
DOI: 10.1007/978-3-319-22053-6_66

web page is referred to other relevant pages, the references on that web page also become important.

The second research field, closely related to the topic of the paper, is a web usage mining (WUM). In terms of WUM, a website's visitor always sends a large amount of information to the server during browsing the website. Most web servers automatically save this information in the form of records stored in log files.

The goal of this paper is to point out the relationship between the estimated importance of web pages (received by the methods of web structure mining) and visitors' actual perception of the importance of individual web pages (obtained by the methods of web usage mining).

We will prove the connection between observed and expected visit rate in the following steps:

- We will summarize results of a pre-experiment where we applied four different approaches to data pre-processing of web server's log file. We applied a sequence rule analysis (the log file collects observed visit rates of individual web pages) with the aim to assess the most suitable steps of data pre-processing for an analysis of website's visitors behaviour.
- We will calculate a PageRank of the individual pages of the examined website. The PageRank represents the probability of accesses to the web pages. In our study, it will represent an expected visit rate of the web pages.
- We will compare the *PageRank* values of the individual web pages and the value of variable s*upport* explaining the actual web page visit rate during the examined period (received from the log file). We will try to prove that the highest dependence of *PageRank* on variable *support* will be in the file where the most suitable steps of data pre-processing for web usage mining were applied (the most suitable steps of data pre-processing were found in the pre-experiment). Hereby we will try to prove the dependence between the estimated and observed probability of accesses to the web pages of the examined website.

We will use possible differences between the expected probability and observed probability of accesses to individual portal web pages for the purpose of identifying suspicious web pages. A suspicious web page is defined as a web page which is not ordered correctly in the hypertext structure of the website.

We will suggest an approach to identifying suspicious web pages based on the comparison of the expected and observed probability of accesses, i.e., we will suggest an approach to determine the web pages where the importance was highly assessed by website's developers but did not achieve the expected real visit rate. Conversely, we will determine the web pages which were underestimated by the developers, but which were frequently visited.

The rest of the paper is structured as follows. The second chapter deals with the related work in the WSM and WUM research area. We describe the tasks related to the initial experiment in the third chapter. The fourth chapter brings the detailed description of the experiment, in which we tried to find dependencies of the value *PageRank* on variable *support*. We provide discussion and conclusions in the last chapter.

2 Related Work

The analysis of users' behaviour represents the main objective of the web usage mining [1, 2]. Data about the accesses of website's visitors is stored in standardized text form of the log files, referred to as Common Log File (CLF).

Data pre-processing refers to the stage of processing of the web server logs for the purpose of identifying meaningful representations. Data cleaning methods are necessary because a WUM is sensitive to noise data. On the other hand, data pre-processing can represent a difficult task when the available data is incomplete or includes erroneous information. According to Cooley, Mobasher, and Srivastava [3] data pre-processing consists of data cleaning (removing irrelevant references and fields, eliminating erroneous references, adding missing references due to caching mechanisms, etc.) and data transformation (user-session identification, path completion [4, 5], etc.).

The web server log file is the primary source of anonymous data about a user (a website's visitor). Anonymous data can also cause a problem with a unique identification of a web page visitor because the visitor can visit the web page repeatedly. Therefore, the web log file can contain multiple sessions of the same visitor.

The objective of this phase of pre-processing is the user session identification [3]. The session identification method using time-window represents the most common method [6]. Using this method, each time we had found subsequent records about the web page requests where the time of the web page displaying had been higher than explicitly selected time, we divided the user visits into several sessions. Explicitly chosen time is denoted as a Standard Time Threshold (STT) and it can take different values: 5 min [7], 10 min [8], 15 min [9], or 30 min [10] or individual threshold [11]. This method is widely used because of its simplicity.

We should briefly mention the alternative methods of the user session identification, which are based on the information stored in cookies files saved on the user's computer. The cookies are tightly bound to the web browser. Even though cookies are considered the most common and the most simple method of user session identification, known issues [12, 13] limit their practical use and have to be replaced by other methods [14].

The main aim of the paper is to use the PageRank (PR) algorithm [15, 16] in WUM domain. Several authors tried to combine WSM, web content mining and WUM methods in several studies. Lorentzen [17] found quite a few studies using a combination of two sub-fields of web log mining.

Usually, the estimation of the web page quality was assured by the PR, HITS or TrustRank algorithms. However, low quality, unreliable data or spam stored in the hypertext structure caused less effective estimation of the web page quality [18, 19]. We can find a review of PR algorithms in Web Mining, their limitations and a new method for indexing web pages in [16]. An interesting approach for using web logs for improvement of website design and organization is described in [20].

Lorentzen [17] noticed that the structure mining is frequently used with other methods. For example, the Markov chain-based Site Rank and Popularity Rank combined structure and usage mining with a co-citation-based algorithm. Another approach used HITS algorithm, semantic clustering, co-link analysis and social network analysis for an automatic generation of hierarchical sitemaps for web sites, or for an automatic exploration of a topical structure of a given academic subject.

Ahmadi-Abkenari [21] introduced a web page importance metric of LogRank that worked based on analysis of different levels of clickstreams in server data set. The importance of each web page was precisely based on the observation period of log data and independent from the downloaded portion of the web.

Agichtein et al. and Meiss et al. [22, 23] used the traffic data to validate the PageRank random surfing model. Su et al. [24] proposed and experimentally evaluated a novel approach for personalized web page ranking and recommendation by integrating an association mining and PageRank.

We also found similar approaches, which combined WUM and WSM methods in other experiments [12, 16, 18, 21, 22, 24–31], but these experiments did not research quality of acquired knowledge about the users' behaviour depending on the selected method of user session identification.

3 Data Pre-Processing

We try to explain, how to link WUM and WSM methods effectively in this section. Firstly, we have to note previously realized experiments, which are not included in this paper, but their results are inevitable for the formulation of findings.

We used data stored in a standard log format in all the experiments described in this paper. We developed a crawler, which went through and analysed web pages. The crawler began on the home page and read all hyperlinks on the examined web page. If the crawler found hyperlinks to the unattended web pages, it added them to the queue. The crawler created a site map which we utilized later in the PR calculation of individual web pages.

At the same time, we used the site map as an input to the path completion algorithm. Besides the site map, the crawler collected information about the level, in which the analysed web page has been in respect to the home page. If the home page was level 1, then all web pages, which had a reference from the home page, would have been level 2. The i^{th} level contained all web pages with references from the i-1 level. It is clear that we considered the highest level of each web page. We considered two categories of the web page levels:

- Category A included web pages of the first and second level, i.e., home page and all web pages, which were accessible from it on one click.
- Category B included all remaining levels.

When the crawler finished, we created a hypertext matrix from the site map. Consequently, we calculated PR for individual web pages according to the formula (1). The value of damping factor d was 0.85.

We used the log files of the university web site. We considered also the session identification method based on cookies in the experiment. It was necessary to change the format of the logs, change the credentials and the manner of writing and reading cookies in the web server. After that, we removed unnecessary records and accesses of crawlers from the log file. The final log file had 573020 records over a period of three weeks. We also removed records, where the information about the cookies was missing (160660 records, 28.04 %). Finally, we obtained the file log with 412360 records.

3.1 Initial Experiment

We prepared an experiment with the aim of verifying the contribution of the proposed method of user session identification using cookies. We used reputable user session identification methodology using STT for this purpose [1, 6, 11, 32, 33]. We applied the proposed method to the four different files with various levels of pre-processing.

At the same time, we intended to find out which of the pre-processed log files was the most suitable for the proposed method. We decided to identify the quality and quantity of the acquired knowledge (behavioural patterns of the users) from the individual log files for this purpose.

We followed the following methodology in the process of examining the influence of data pre-processing on the quality and quantity of extracted knowledge [33]:

1. Data acquisition – definition of observed variables in the log file (IP address, access date and time, URL).
2. Data matrices creation from the log file (information about users' accesses) and the site map (information about the web content).
3. Data pre-processing on the different levels.
4. Data analysis – user behavioural pattern finding in individual files. We used the Apriori algorithm for extraction of sequence rules implemented in the Sequence Association and Link Analysis [34] Module of STATISTICA.
5. Output data understanding – the creation of data files from the outputs of the analysis of individual files and basic characteristics calculation.
6. Comparison of obtained knowledge from the files, which were pre-processed at the different levels. We evaluated the acquired knowledge in terms of the quality and quantity of found sequence rules – user behavioural patterns. We took great care in:

- Comparison of proportion of found rules in examined files.
- Comparison of proportion of useful, trivial, or inexplicable rules in examined files.
- Comparison of variables *support* and *confidence* of found rules in examined files.

We prepared data at some levels. We obtained the final set of files:

- File **A1** – session identification using STT without the path completion,
- File **A2** – session identification using STT with the path completion,
- File **B1** – session identification using cookies without the path completion,
- File **B2** – session identification using cookies with the path completion.

We used STT = 10 min in the cases of files A1 and A2 and 10 min for cookies expiration in the case of files B1 and B2.

3.2 Results of Initial Experiment

We examined users' accesses to the web site of the university during three weeks. We obtained the sequence rules from the frequented sequences, which accomplished the minimal support (min $s = 0.005$) as a result of the analysis (Table 1). We obtained frequented sequences previously from the identified sequences, i.e., from the visits of individual users in the observed period.

Table 1. Discovered sequence rules in individual files.

Body	->	Head	A1	A2	B1	B2
(/), (/admissions)	->	(/admissions/admissions-results)	1	1	1	1
…	->	…	…	…	…	…
(/university-structure)	->	(/university-structure)	0	1	0	1
(/university-structure)	->	(/university-structure/faculty-of-natural-sciences)	1	1	1	1
…	->	…	…	…	…	…
(/study/accredited-study-programs)	->	(/study)	0	1	0	1
Count of derived sequence rules			51	197	43	227
Percent of derived sequence rules (Percent 1's)			21.52	83.12	18.14	95.78
Percent 0's			78.48	16.88	81.86	4.22
Cochran Q test			Q = 443.3120, df = 3, p < 0.000000			

We could see the high consistency (compliance) between the results of the sequence rule analysis in terms of the portion of the found rules in the files without the path completion (A1, B1). Simultaneously, we could see the similar compliance in the case of files with a path completion (A2, B2).

We extracted most of the rules from the file with the identified user sessions and completed paths. More rules were discovered in the files with the path completion (A2, B2).

The assessment of the quality of obtained sequence rules represented the next step in results evaluation. We assessed two characteristics - *support* and *confidence*. We found differences in quality and quantity of the found sequence rules between individual files regarding the values of variable *support*. Statistically, significant differences were found between the files without path completion (A1, B1) and between the files with path completion (A2, B2).

We should have considered files with path completion (A2, B2) as the best-pre-processed files for the extraction of user behavioural patterns.

4 Finding Dependences Between Variables *PageRank* and Variable *Support*

We merged WSM methods (PageRank) with WUM methods (sequence rule analysis) in the following experiment. We tried to answer the question if the expected visit rate of individual web pages (calculated using PR) correlates with the observed visit rate of the web pages, which were found by the WUM method in the previously described experiment and expressed by the value of variable *support*.

Variable *support* is defined as *support*(X) = $P(X)$. In other words, item X has a support s if s % of transactions contain X, i.e. the variable *support* means the frequency of occurrence of given set of items in the database. It represents the probability of visiting a particular web page in identified sequences (sessions). We assumed that the data reliability used in WUM would be increasing with the growth of dependence between values of *PR* and *support*.

4.1 Results

The value of *PR* and the level of the web page were added to the data obtained from the log file. It means that we calculated PR and assigned the appropriate level to each record of the log file. We made these changes to the log file examined in the initial experiment (Sect. 3.1). We analysed not only this log file, but also the log files which have been pre-processed in the same manner.

The variable *support* means the probability of individual web page visits in identified sessions. We examined the variable *support* from available statistics, and subsequently we analysed only the web pages with minimal support 0.5 %.

Table 2 shows the dependence of *PR* on the variable *support* calculated from the files with different level of data pre-processing. A directly proportional relationship was identified in all examined files. There were evident variations from normality. Therefore, we used non-parametric correlation [35] for calculation of dependence rate between *PR* and *support*.

Table 2. The dependence between *PR* and variable *support* in examined files.

Total	Valid N	Spearman R	t(N−2)	p-level
PageRank & support (A1)	47	0.4052	2.972739	0.004728
PageRank & support (A2)	42	0.4248	2.967526	0.005049
PageRank & support (B1)	46	0.4004	2.898073	0.005834
PageRank & support (B2)	39	0.4687	3.227136	0.002619

We identified the medium dependence of *PR* on variable *support* (Table 2). The dependence was greater in the files with identified paths. The correlation coefficients were statistically significant at the 1 % significance level. The greatest dependence was reached in file B2. Following these findings we could assume that the method of session identification using cookies in conjunction with the path completion had the greatest impact on the data reliability.

If we considered the web structure, i.e. the position of the web page in web site, we obtained similar results. We regarded two categories of web pages:

- Category A contains web pages of the first and second level.
- Category B contains remaining levels.

The impact of data pre-processing was not significant in category A. The coefficients of correlation (Table 3) were not significant (Spearman R < 0.4; p > 0.05).

On the opposite, we could suppose the path completion had an impact on the data reliability (Spearman R > 0.4; p < 0.05) in the case of category B (Table 4).

Table 3. The dependence between *PR* and variable *support* in category A.

Category A	Valid N	Spearman R	t(N−2)	p-level
PageRank & support (A1)	19	0.3809	1.698370	0.107664
PageRank & support (A2)	18	0.3635	1.560512	0.138198
PageRank & support (B1)	19	0.3668	1.625806	0.122384
PageRank & support (B2)	17	0.3936	1.658376	0.118001

Table 4. The dependence between *PR* and variable *support* in category B.

Category B	Valid N	Spearman R	t(N−2)	p-level
PageRank & support (A1)	28	0.3710	2.037140	0.051936
PageRank & support (A2)	24	0.4476	2.347991	0.028276
PageRank & support (B1)	27	0.3568	1.909797	0.067698
PageRank & support (B2)	22	0.4245	2.096915	0.048918

We also compared the values of *PR* and variable *support* for given categories. We considered the position of the web page in the structure of the web site. The distribution of observed variables was asymmetric. Therefore, we used non-parametric Mann-Whitney U Test for differences testing.

We found statistically significant difference of *PR* values in categories A and B at the 0.1 % significance level from the results of Mann-Whitney U Test (Table 5). The value of median was 10.3 and middle 50 % values were from the interval 9.4–10.6 in the category A. The value of median was 0.5 and middle 50 % values were from the interval 0.2–1.2 in the category B.

Table 5. Testing differences: *PR* x *category*.

	Rank sum A	Rank sum B	U	Z	p-level
PageRank	664.5	463.5	57.5	4.519811	0.000006

The statistically significant differences in variable *support* between categories A and B in examined files (Table 6) were not proven (p > 0.05).

Table 6. Testing differences: variable *support* x *category*.

	Rank sum A	Rank sum B	U	Z	p-level
support (A1)	534.0	594.0	188.0	1.690864	0.090864
support (A2)	448.5	454.5	154.5	1.563110	0.118028
support (B1)	520.0	561.0	183.0	1.639722	0.101064
support (B2)	402.0	378.0	125.0	1.755968	0.079095

The median of variable *support (A1)* was 1.6 and the middle 50 % of values were from the interval 0.6–5.2 in the category A. In contrast to these values, the median was 0.9 and the middle 50 % of values were from the interval 0.6–1.5 in the category B. This means the values of *support* were more homogeneous in category B than in category A from the variability point of view.

We also achieved similar results in files A2, B1 and B2. We noticed the differences in the variable *support* only in the variability between categories A and B.

5 Discussion and Conclusion

The authors of the original idea of PR introduced PR as a probability that the random visitor accessed a particular web page. The log file and WUM methods represented the observed visit rate of individual web pages in realized experiment. We proved experimentally that the found expected visit rate (*PR*) correlate with the observed visit rate expressed as a value of the variable *support*.

As we noted previously, we have to modify the log file in the pre-processing phase in order to obtain the real user sessions. The quality of acquired knowledge about user's behaviour depended on the selected method of the session identification and executed changes.

We proved in the experiment that there was a higher dependence between *PR* and the variable *support* in the visit rate of the examined web pages in well-prepared files with identified user sessions.

We found out statistically significant differences between *PR* and variable *support* in the files with the application of more advanced methods of session identification. We obtained the same results in the initial experiment (Sects. 3.1 and 3.2), where we tried to compare session identification methods in the four files with different levels of data pre-processing. We proved the results of the initial experiment by using a different methodology.

We verified a new proposed methodology for the comparison of methods of data pre-processing in WUM in terms of the reliability of the obtained data. We could follow these steps:

1. Data acquisition.
2. Creating of data matrices from the log file and site map.
3. Data pre-processing at different levels.
4. PR calculation for individual web pages.
5. Variable *support* calculation of the web page selected from the log files and assignment of PR to the individual web pages. In this case, we considered only web pages with value of *support* > 0.5 %.
6. Data understanding and creation of data files from the calculated characteristics.
7. Comparison of obtained characteristics in term of the dependence of *PR* on variable *support* from log files, which were pre-processed at different levels.

We omitted some steps in contrast to the method introduced in the Sect. 3.2. We did not extract the rules and did not analyse the extracted rules in terms of their quantity and quality. We verified suitability of these steps in terms of the dependence on *PR* and variable *support*. We focused on the suitability verification of proposed steps from the point of view of the dependence between *PR* and variable *support*.

The proposed methodology did not involve the extraction of sequence rules from examined files. It only claimed that well-prepared files are files which best reflect the dependence of the expected and observed visit rate (expected and the real probability).

On the other hand, the usage analysis involved rules extraction. It means that the proposed methodology is only an alternative or supplementary method. It serves mainly for experimental purposes and results verification of the original methodology (used in Sect. 3.1).

Acknowledgments. 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 and project KEGA 015UKF-4/2013 Modern computer science – New methods and forms for effective education.

References

1. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explorations Newsletter 1, pp. 12–23 (2000)
2. Romero, C., Ventura, S., Zafra, A., Bra, P.D.: Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. Comput. Educ. **53**, 828–840 (2009)
3. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowl. Inf. Syst. **1**, 5–32 (1999)
4. Zhang, C., Zhuang, L.: New path filling method on data preprocessing in web mining. In: Proceedings of Computer and Information Science 1, pp. 112–115 (2008)
5. Li, Y., Feng, B., Mao, Q.: Research on path completion technique in web usage mining. In: Proceedings of the 2008 International Symposium on Computer Science and Computational Technology, vol. 01, pp. 554–559. IEEE Computer Society (2008)
6. Huynh, T., Miller, J.: Empirical observations on the session timeout threshold. Inf. Process. Manag. **45**, 513–528 (2009)
7. Downey, D., Dumais, S., Horvitz, E.: Models of searching and browsing: languages, studies, and applications. In: Proceedings of the 20th international joint conference on Artifical intelligence, pp. 2740–2747. Morgan Kaufmann Publishers Inc., Hyderabad (2007)
8. Chien, S., Immorlica, N.: Semantic similarity between search engine queries using temporal correlation. In: Proceedings of the 14th International Conference on World Wide Web, pp. 2–11. ACM, Chiba, (2005)
9. He, D., Göker, A.: Detecting session boundaries from web user logs. In: Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research, pp. 57–66 (2000)
10. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge discovery in Data Mining, pp. 239–248. ACM, Chicago (2005)
11. Mehrzadi, D., Feitelson, D.G.: On extracting session data from activity logs. In: Proceedings of the 5th Annual International Systems and Storage Conference, pp. 1–7. ACM, Haifa (2012)
12. Guerbas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., Ridley, M., Alhajj, R.: Effective web log mining and online navigational pattern prediction. Knowl. Based Syst. **49**, 50–62 (2013)

13. Cooley, R.: Web usage mining: discovery and application of interesting patterns from web data. Ph.D. thesis. University of Minnesota (2000)
14. Schmitt, E., Manning, H., Paul, Y., Tong, J.: Measuring Web Success. Forrester Report (1999)
15. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. **30**, 107–117 (1998)
16. Jain, A., Sharma, R., Dixit, G., Tomar, V.: Page ranking algorithms in web mining, limitations of existing methods and a new method for indexing web pages. In: Proceedings of the 2013 International Conference on Communication Systems and Network Technologies, pp. 640–645. IEEE Computer Society (2013)
17. Lorentzen, D.G.: Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. Scientometrics **99**, 409–445 (2014)
18. Lili, Y., Yingbin, W., Zhanji, G., Yizhuo, C.: Research on PageRank and hyperlink-induced topic search in web structure mining. In: Conference Research on PageRank and Hyperlink-Induced Topic Search in Web Structure Mining, pp. 1–4 (2011)
19. Wu, G., Wei, Y.: Arnoldi versus GMRES for computing PageRank: a theoretical contribution to google's PageRank problem. ACM Trans. Inf. Syst. **28**, 1–28 (2010)
20. Xu, G., Zhang, Y., Li, L.: Web Mining and Social Networking Techniques and Applications. Springer, Heidelberg (2011)
21. Ahmadi-Abkenari, F., Selamat, A.: A clickstream based web page importance metric for customized search engines. In: Nguyen, N. (ed.) Transactions on Computational Collective Intelligence XII, vol. 8240, pp. 21–41. Springer, Berlin Heidelberg (2013)
22. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research And Development in Information Retrieval, pp. 19–26. ACM, Seattle (2006)
23. Meiss, M.R., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: Ranking web sites with real user traffic. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 65–76. ACM, Palo Alto (2008)
24. Su, J.-H., Wang, B.-W., Tseng, V.S.: Effective Ranking and Recommendation on web page retrieval by integrating association mining and PageRank. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, pp. 455–458. IEEE Computer Society (2008)
25. Srikant, R., Yang, Y.: Mining web logs to improve website organization. In: Proceedings of the 10th International Conference on World Wide Web, pp. 430–437. ACM, Hong Kong (2001)
26. Liu, H., Keselj, V.: Combined mining of web server logs and web contents for classi fying user navigation patterns and predicting users' future requests. Data Knowl. Eng. **61**, 304–330 (2007)
27. Das, R., Turkoglu, I.: Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Syst. Appl. **36**, 6635–6644 (2009)
28. Yang, Q., Ling, C., Gao, J.: Mining web logs for actionable knowledge. In: Liu, J., Zhong, N. (eds.) Intelligent Technologies for Information Analysis, pp. 169–191. Springer, Heidelberg (2004)
29. Eirinaki, M., Vazirgiannis, M.: Usage-based PageRank for web personalization. In: Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 130–137. IEEE Computer Society (2005)
30. Masseglia, F., Poncelet, P., Teisseire, M.: Using data mining techniques on web access logs to dynamically improve hypertext structure. SIGWEB Newsletter, 8, pp. 13–19 (1999)

31. Tripathy, A., Patra, P.K.: A Web mining architectural model of distributed crawler for internet searches using PageRank algorithm. In: Proceedings of the 2008 IEEE Asia-Pacific Services Computing Conference, pp. 513–518. IEEE Computer Society (2008)
32. Fang, Y., Huang, Z.: An improved algorithm for session identification on web log. In: Wang, F., Gong, Z., Luo, X., Lei, J. (eds.) Web Information Systems and Mining, vol. 6318, pp. 53–60. Springer, Heidelberg (2010)
33. Munk, M., Kapusta, J., Švec, P.: Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor. Procedia Comput. Sci. **1**, 2273–2280 (2010)
34. 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, pp. 487–499. Morgan Kaufmann Publishers Inc., (1994)
35. 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), pp. 329–338 (2013)