Analysis of Differences between Expected and Observed Probability of Accesses to Web Pages

Jozef Kapusta, Michal Munk, Martin Drlík

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

Abstract. The paper introduces an alternative method for website analysis that combines two web mining research fields - discovering of web users' behaviour patterns as well as discovering knowledge from the website structure. The main objective of the paper is to identify the web pages, in which the value of importance of these web pages, estimated by the website developers, does not correspond to the actual perception of these web pages by the visitors. The paper presents a case study, which used the proposed method of the identification suspicious web pages using the analysis of expected and observed probabilities of accesses to the web pages. The expected probabilities were calculated using the PageRank method and observed probabilities were obtained from the web server log file. The observed and expected data were compared using the residual analysis. The obtained results can be successfully used for the identification of potential problems with the structure of the observed website.

Keywords: web usage mining, web structure mining, PageRank, support, observed, visit rate, expected visit rate.

1 Introduction

The aim of the web portal designers or developers is to provide information to users in a clear and understandable form. Information on web pages is interconnected by the hyperlinks. The website developer of designer affects visitors' behaviour by references. He/she indicates the importance of information displayed on web pages through the references. Probably more references head to more important web pages, which are directly accessed 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 the opposite direction. The website providers can obtain the amount of information about their users or about users' behaviours, needs or interests.

The goal of this paper is to point out the connection between the estimated importance of web pages (obtained by methods of web structure mining) and visitors' actual perception of the importance of individual pages (obtained by methods of web usage mining – a part of web log mining). We can use possible differences between the expected probability and observed probability of accesses to the individual portal

web pages to identify suspicious pages. Suspicious pages are defined as pages that are not ordered correctly in the hypertext portal structure.

By knowledge discovery from web page structure (web structure mining, WSM) [1], we focused on the analysis of the quality and importance of web pages based on references among web pages. Determination of web page importance is based on the idea that the degree to which we can rely on the web page quality is transferred by references among web pages. If the web page is referred to by other important pages, the references on that page also become important. By the web usage mining (WUM), we start from the fact, that a user usually posts a large amount of information to the server during his visit of the web page. The most web servers automatically save this information in the form of log file records.

2 Related work

The authors tried to combine web structure mining, web content mining and web usage mining methods in several studies. The combination of methods and techniques of these research fields could help to solve some typical types of web structure analysis issues.

The authors of similar experiments examined several other methods of web page quality estimation, which plays a crucial role in the contemporary web searching engines.

Usually, the estimation of web page quality was assured by the PR or TrustRank algorithms. However, low quality, unreliable data or spam stored in the hypertext structure caused less effective estimation of the web page quality.

Liu et al. [2] further utilized learning algorithms for web page quality estimation based on the content factors of examined web pages (the web page length, the count of referred hyperlinks).

Chua and Chan [3] dealt with the analysis of selected properties of examined web pages. They combined the content, structure and character of hyperlinks of web pages for the web page classification. They wanted to define web page classifiers for thematically oriented search engines more precisely. Each analysed web page was represented by a set of properties related to its content, structure and hyperlinks. Finally, they stated 14 properties of web pages, which they used consequently as the input of machine learning algorithms.

Jacob et al. [4] designed an algorithm WITCH (Webspam Identification Through Content and Hyperlinks), which combined a web structure and web content mining methods for the purpose of spam detection. We also found a similar approach in other experiments [5-7].

Lorentzen [8] found only a few studies, which combine two sub-fields of web mining. The most of the research has focused on a combination of the usage and content mining methods, but he also mentioned some examples of structure mining, which could be said to be web mining's equivalent to link analysis. For example, the Markov chain-based Site Rank and Popularity Rank combine structure and usage mining with a co-citation-based algorithm for the automatic generation of hierarchical

sitemaps for websites, or for the automatic exploration of the topical structure of a given academic subject, based on the HITS algorithm, semantic clustering, co-link analysis and social network analysis.

Usually, as we wrote previously, 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 [9, 10]. Jain at al. [11] provided a detailed review of PR algorithms in Web Mining, their limitations and a new method for indexing web pages.

Ahmadi-Abkenari [12] introduced web page importance metric of LogRank that works based on analysis on server level clickstream data set. The application of this metric means the importance of each page is based on the observation period of log data and independent from the downloaded portion of the Web. Agichtein et al. and Meiss et al. [13, 14] used the traffic data to validate the PageRank random surfing model. Su et al. [15] proposed and experimentally evaluated a novel approach for personalized page ranking and recommendation by integrating association mining and PageRank.

3 Data Pre-processing

We mentioned previously, that we tried to connect web usage mining and web structure mining methods for the purpose of identification the differences between expected and observed accesses to the web pages. We describe the details of the proposed approach in this section.

3.1 Data Pre-processing for PR Calculation

We developed the 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 had created a site map which we have utilized later in the PR calculation of individual pages. The web crawler implemented the method, which has been operating in several steps:

- URL selection from the queue.
- An analysis of the content of selected web page for the purpose of finding new URL references.
- New URL references added to the queue.

The crawler was simple because it scanned only the hyperlinks between web pages. We consider this as the main limitation of the proposed method, because the crawler did not regard the actual position of the hyperlinks within the web page layout, which has a strong influence on the probability of being accessed by a website visitor.

Consequently, we calculated PR for different web pages. Brin and Page [16], the authors of the PR, introduced the Random Surfer Model. This model assumed that the

user had clicked on the hyperlinks, he/she had never returned, and they have started on other random web pages.

PR of the web page *i* (*PR(i)*) is defined as

$$PR(i) = (1 - d) + d \sum_{(j,i) \in E} \frac{PR(j)}{O_j} \quad (1)$$

where *d* is Dumping Factor (*0≤d<1*), *E* is a set of oriented edges, i.e., hyperlinks between web pages, *j* is a web page with a reference on the web page *i*, and *Oj* is the count of hyperlinks referred to the other web pages from the web page *j*.

Therefore, the authors proposed Dumping Factor *d* ⟨0, 1). Dumping factor represented the probability that the random surfer would continue on the next web page. The value 1- *d* meant the probability that the user would start on the new web page. The typical value of the variable *d* is usually 0.85 [17]. We can iterate this calculation until the value of *Pr(i)* begins to converge to the limit value [16].

3.2 Data Pre-processing for Visit Rate Finding of Individual Web Pages

We used the log files of the university web site. 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 prepared data at some levels:

- Data with session identification using standard time threshold (STT) [18-20]. The session identification method using standard time threshold (time-window) represents the most common method. 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. We used STT = 10 minutes.
- Data with path completion [21, 22]. The reconstruction of the web site visitors' activities represents another issue for WUM. This technique does not belong to the session identification approaches. It is usually the next step of the data preprocessing phase [23, 24]. The main aim of this stage is to identify any significant accesses to the web page which are not recorded in the log file.
  For example, if the user returns to the web page during the same session, the second attempt will probably display the cached version of the web page from the Internet browser. The path completion technique provides an acceptable solution to these problems. It assumes that we can add the missing records to the web server log file using the site map or eventually using the value of the variable *referrer* stored in the log file [25, 26].

The observed visit rate of the web pages were found by the WUM method and represented by the value of variable *support*. The variable *support* is defined as *support(X) = P(X)*. In other words, item *X* has 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).

4 Analysis of Differences between Expected and Observed Probability of Accesses to Web Pages

We could see input data of web mining from two different views:

- Data, which depends on the web page developers, this data represents the input of the web structure mining.
- Data, which depends on the web page visitors, this data represents the input of the web usage mining.

Of course, we could dispute that the visitors' behaviour was determined by the structure and the content of the web page and vice versa, but we have had the primary origin of data in mind.

These two groups of data are interdependent. The web page developers should create web pages that reflect the needs of their visitors. We proved the dependence of values PR and real visit rate expressed as values of the variable *support*.

We analysed the dependence of expected values of the web page accesses on real values. The value of PR for individual web pages represented the probability of being visited by a random visitor. At the same time, PR expressed the importance of the web page from the web developer's perspective. If the web page had been important, the developer created more references directed to this web page than to other, less important, web pages. We compared the importance of the web page given by the developer with the real importance of this web page from the visitor's point of view.

We will describe one method of finding the differences between the expected and observed probability of accesses to the individual web pages of the website and analyse the obtained results.

5 Residual analysis

We supposed that we had a log file from a web server with users' sessions identified using the STT method. Simultaneously, we calculated PR for each web page on the examined web site. Finally, we calculated values of the variable *support*. We only used the web pages with a value of variable *support* greater than 0.5 in the residual analysis.

We considered the comparable values for the comparison of expected (PR) and observed (support) visit rate. The variable *support* was from the interval 0-100 and represented the probability of visiting a particular web page in the identified sessions.

We should be aware that the values of PR of individual web pages created the probability distribution of the visits together. Therefore, the sum of PR should be 1. We transformed it into relative values for that reason.

We found inspiration in the residual analysis. The main idea of this method assumes that

Data = prediction using model (function) + residual value.

If we subtracted the values obtained from the model (expected values) from the observed values, we would have got errors (residual values). We could analyse the residual values for the purpose of the model appraisal.

The selection residues e_i are defined as

$$e_i = y_i - \hat{y}_i \ , \tag{2}$$

where $\hat{y}_i$ are expected values predicted by the model and y_i are observed values.

The residual analysis serves for the purpose of the model validity verifying and its improving because it helps to find out the relationships, which the model did not consider. For example, we can use the residual analysis for the regression model stability verifying, i.e., we can identify the incorrectness of the selected model using the correlated chart of residues and independent variable.

The values of variable *support* represented the observed values in the described experiment. The values of the variable PR represented the expected values. As we mentioned earlier, the main objective of the residual analysis was the identification of the outliers. We could visualize the residues in the charts of defined cases (Fig. 1).

Case		Residual -5,81 6,49	Residual
1	(/university-structure/students-dormitories)		-5.4575926
2	(/about-the-university)		-4.6926729
3	(/study)		0.5355837
4	(/university-structure)		-1.1946187
5	(/about-the-university/contacts-list)		3.1038457
6	(/admissions)		5.9986803
	...		...
	...		...
44	(/direct-links/925-academic-year-schedule)		0.559619
45	(/documentsy/university-structure/students-dormitories/04062013.pdf)		0.628261
	Min.		-5,8135952
	Max.		6,4851526
	Mean		-0,1935154
	Median		0,4314872

Fig. 1. Chart of residues example

The chart visualizes calculated values of residues, minimum, maximum, mean and median for each case. We did not consider the home page of the website in this case study because the value of the variable *support* for this page was equal to 53.51. The average value of the variable *support* was 3.09. Therefore, the value of the main page would have distorted the residues of other web pages in the charts.

We created a chart with expected, observed and residual values for better understanding (Fig. 2). The individual web pages ordered by the PR represent the x axis. We can see the web page identifier in the chart of residues. We could see from this chart that the expected values of PR and observed values of the variable *support* were different. The residuum had to be equal zero in the ideal case, i.e., the expected and observed values should be the same. It implies, the structure of the references to a

given web page created by the web developers would be better if the value of the residuum is closer to the x axis.

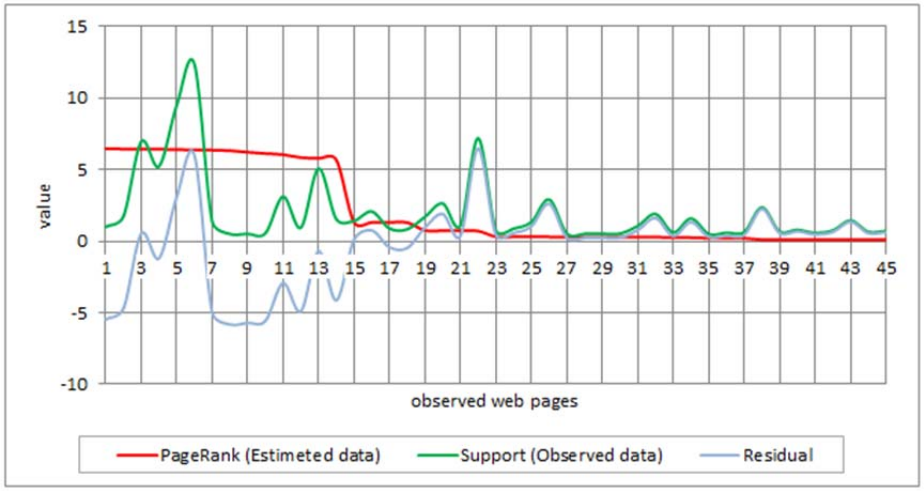

Fig. 2. Chart with expected, observed and residual values

6 Residual Outliers Identification

The identification of outliers is the objective of the residual analysis. The outlier identifies potentially "suspicious" web pages. In this case, the outliers identified the web pages where the structure of the web site (the intention of the developers, creators) did not reflect the real behaviour of the visitors. We created the chart of residues for the outliers' identification (Fig. 3). Considering the theory of residues we could have identified the outliers using rule $\pm 2\sigma$. It means that we considered the cases which were out of the interval

Average of differences ± *2 standard deviation of differences*.

We calculated the following boundary values for the selected web pages of the website: -5.696351; 5.30932. The chart (Fig. 3) visualizes the identification of outliers. As we can see, we identified four "suspicious" web pages.

We identified two web pages with the residuum greater than $+2\sigma$, which were underestimated by web developers. Even though the web pages had few references from other web pages, the visit rate of these web pages was high. It could have been caused by the seasonal importance of the web pages' content. On the other hand, it is clear that these web pages should have had references from more relevant web pages on the web site.

The main menu of the examined web site caused the main problem of the described experiment. The main menu was available on each web page of the web site. It means that there were always the direct references to the main parts of the web site from all

web pages. Therefore, there occurred the evident difference between the value of PR of web pages available directly from the main menu and other web pages.

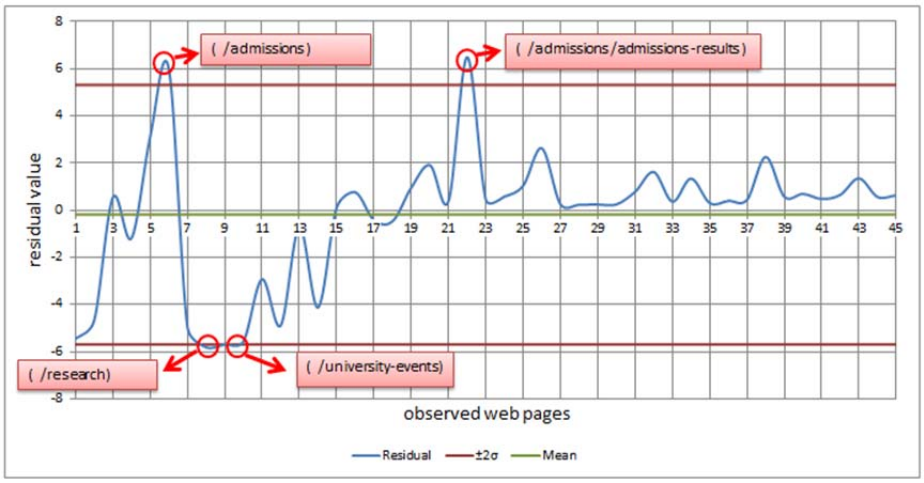

Fig. 3. Visualization the identification of outliers

At the same time, we identified two "suspicious" web pages with the residuum greater than -2σ. These web pages were overestimated by the web site developers. Even though the web pages had many references from other web pages, these web pages obtained a small visit rate.

7 Discussion

We identified the suspicious web pages on the basis of the rule ±2σ. We analysed 45 web pages and found four suspicious pages. It is questionable if the number of the suspicious web pages is sufficient. If we identify a greater number of suspicious web pages we could assess the boundary values using the quartile interval$(Q_I - 1{,}5Q; Q_{III} + 1{,}5Q)$. Figure 4 (Fig. 4) depicts the visualization of the suspicious web pages using the mentioned boundary values.

We utilize the results of the residual analysis to recommend website structure changes. In the described case, we should create references (hyperlinks) on the main page or add other items to the main menu for all pages where the value of residuum was greater than the boundary value. On the other hand, we should change the structure of the references on web pages where the value of residuum was lower than the boundary value.

Web usage mining methods examine the behaviour of visitors on the website. In general, we obtained the set of rules using these methods. We could have evaluated each useful rule subjectively if the rule were in accordance with the idea of the website developers.

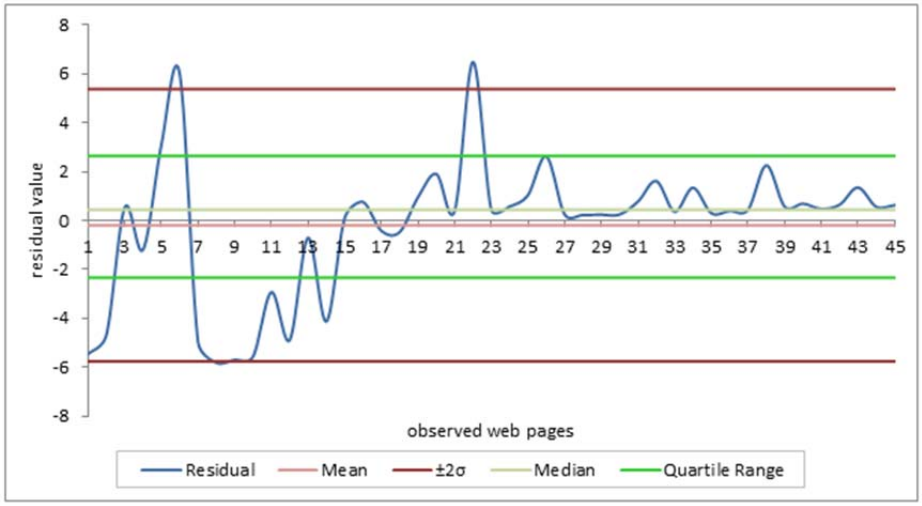

Fig. 4. Visualization boundary values of residual analysis

Consequently, we might identify problematic parts of the examined web site. We might compare the intention of the web site developers with the real visits of the web-site using a combination of web structure and web usage mining methods. Moreover, we might emphasize the potential differences using residual analysis.

The proposed method also has some limitations. We have already mentioned the limited behaviour of the crawler, which was used for the PR calculation of individual pages. It is necessary to take into account also other characteristics of the hyperlinks to improve the obtained results, i.e., their position on a given web page, detection of unwanted clicks, etc.

8 Conclusions

We paid attention to the new web structure mining method in this paper. We chose the algorithm for PR calculation for estimations of importance and quality of the individual web page. The quality of a given web page depends on the number and quality of the web pages which refer to it. We selected the PR method because this method expresses the probability of visiting of given web page by a random visitor.

We presented the case study of the proposed method of identification "suspicious" web pages in the last chapter. Following the conclusions of the previous experiments, we assumed that the expected visit rate would correlate with the real visit-rate.

We utilized the potential advantages of joining web structure and web usage mining methods in the residual analysis. We tried to identify the potential problems with the structure of the web site. Whereas the sequence rules analysis can only uncover the potential problems of web pages with higher visit rate, the proposed method of residual analysis can also detect the web pages with a low visit rate.

9 Acknowledgements

This paper is published with the financial support of the project of Scientific Grant Agency (VEGA), project number VEGA 1/0392/13 and Cultural and Educational Grant Agency (KEGA), project number 067UKF-4/2012.

10 Literature

1. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl. 1, 12-23 (2000)
2. Liu, Y., Zhang, M., Cen, R., Ru, L., Ma, S.: Data cleansing for web information retrieval using query independent features. Journal of the American Society for Information Science and Technology 58, 1884-1898 (2007)
3. Chau, M., Chen, H.: A machine learning approach to web page filtering using content and structure analysis. Decision Support Systems 44, 482-494 (2008)
4. Jacob, A., Olivier, C., Carlos, C.: WITCH: a new approach to Web spam detection. Yahoo! Research Report No. YR-2008-001 (2008)
5. Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F.: Know your neighbors: web spam detection using the web topology. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 423-430. ACM, Amsterdam (2007)
6. Gan, Q., Suel, T.: Improving web spam classifiers using link structure. Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, pp. 17-20. ACM, Banff (2007)
7. Ntoulas, A., Najork, M., Manasse, M., Fetterly, D.: Detecting spam web pages through content analysis. Proceedings of the 15th international conference on World Wide Web (WWW), pp. 83-92, Edinburgh (2006)
8. Lorentzen, D.G.: Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. Scientometrics 99, 409-445 (2014)
9. 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)
10. 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)
11. 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. Proceedings of the 2013 International Conference on Communication Systems and Network Technologies, pp. 640-645. IEEE Computer Society (2013)
12. 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)
13. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19-26. ACM, Seattle, Washington, USA (2006)

14. Meiss, M.R., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: Ranking web sites with real user traffic. Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 65-76. ACM, Palo Alto, California, USA (2008)
15. Su, J.-H., Wang, B.-W., Tseng, V.S.: Effective Ranking and Recommendation on Web Page Retrieval by Integrating Association Mining and PageRank. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03, pp. 455-458. IEEE Computer Society (2008)
16. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks 107-117 (1998)
17. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report. Technical report. Standford Digital, Standford (1998)
18. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information System 1, (1999)
19. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide Web. Comput. Netw. ISDN Syst. 27, 1065-1073 (1995)
20. Pirolli, P., Pitkow, J., Rao, R.: Silk from a sow's ear: Extracting usable structures from the Web. In: Conference Silk from a sow's ear: Extracting usable structures from the Web. (1996)
21. Dhawan, S., Lathwal, M.: Study of Preprocessing Methods in Web Server Logs. International Journal of Advanced Research in Computer Science and Software Engineering 3, 430-433 (2013)
22. Li, Y., Feng, B., Mao, Q.: Research on Path Completion Technique in Web Usage Mining. Proceedings of the 2008 International Symposium on Computer Science and Computational Technology - Volume 01, pp. 554-559. IEEE Computer Society (2008)
23. Gong, W., Baohui, T.: A New Path Filling Method on Data Preprocessing in Web Mining. In: Conference A New Path Filling Method on Data Preprocessing in Web Mining, pp. 1033-1035. (2012)
24. Klocoková, D.: Integration of heuristics elements in the web-based environment: Experimental evaluation and usage analysis. Procedia - Social and Behavioral Sciences 15, 1010-1014 (2011)
25. Chitraa, V., Davamani, A.S.: An Efficient Path Completion Technique for web log mining. IEEE International Conference on Computational Intelligence and Computing Research, (2010)
26. Zhang, C., Zhuang, L.: New Path Filling Method on Data Preprocessing in Web Mining. Proceedings of Computer and Information Science 1, 112-115 (2008)