Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia Computer Science 18 (2013) 1747 – 1756
Procedia Computer Science

2013 International Conference on Computational Science

Pillar 3 and Modelling of Stakeholders' Behaviour at the Commercial Bank Website during the Recent Financial Crisis

Michal Munk[a], Anna Pilkova[b], Jozef Kapusta[a], Peter Svec[a]*, Martin Drlik[a]

[a]Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia
[b]*Commenius Univeristy in Bratislava, Šafárikovo nám. 6, 818 06 Bratislava, Slovakia*

Abstract

The paper analyses domestic and foreign market participants' interests in mandatory Basel 2, Pillar 3 information disclosure of a commercial bank during the recent financial crisis. The authors try to ascertain whether the purposes of Basel 2 regulations under the Pillar 3 - Market discipline, publishing the financial and risk related information, have been fulfilled. Therefore, the paper focuses on modelling of visitors' behaviour at the commercial bank website where information according to Basel 2 is available. The authors present a detailed analysis of the user log data stored by web servers. The analysis can help better understand the rate of use of the mandatory and optional Pillar 3 information disclosure web pages at the commercial bank website in the recent financial crisis in Slovakia. The authors used association rule analysis to identify the association among content categories of the website. The results show that there is in general a small interest of stakeholders in mandating the commercial bank´s disclosure of financial information. Foreign website visitors were more concerned about information disclosure according to Pillar 3, Basel 2 regulation, and they have less interest in general information about the bank than domestic ones.

Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science

Keywords: Association rules; web log miming; risk management; business intelligence; financial regulation; market discipline; commercial banks; statistical analysis;

1. Introduction

Market discipline (MD) has been recognized as an important supplement of regulators´ supervisory efforts in banking. Its concept and related issues have been studied from different perspectives, and many findings have been presented also in relation to the last financial crisis. There are numerous definitions of market

* Corresponding author. Tel.: +421-37-6408-675; fax:.+421-37-6408-555.
E-mail address: psvec@ukf.sk.

Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science
doi:10.1016/j.procs.2013.05.343

discipline (MD) in the literature. We define MD in banking as the mechanism by which market participants monitor and discipline excessive risk-taking behaviour by banks [1]. We agree with those authors who have pointed out that MD is less about the market and more about the institutional framework – information, incentives and control – used to reduce the problems of moral hazard and asymmetric information. In our research we focus on the issue of market stakeholders' behaviour in relation to information disclosure according to Pillar 3, Basel 2. It is well known that the importance of MD has been formally stressed through the latest architecture of the banking regulatory framework Basel 2 and 3 [2-3]. Within these frameworks market discipline is one of the three pillars (Pillar3), along with capital regulation (Pillar 1) and supervision (Pillar2). Even Basel II documents did not define market discipline, but it mandates different information, which must be disclosed by banks. Those are quantitative and qualitative disclosure requirements regarding capital, risk exposures, risk management practices and capital adequacy.

The key idea behind of Pillar 3 is that based upon disclosed information market participants should be able to understand and subsequently judge the relevance of the bank risk management and try to "discipline" risky banks by requiring higher spreads for deposits or even by refusing new funding for these banks. In general, the disclosures are to be made to and for the benefit of the market. Neither Basel Committee nor European Union related legislation [4-5] asks for a standardized format of information disclosure, and frequency, is limited to a minimum of one time per year. The National Bank of Slovakia (NBS) issued two decrees [6-7] that standardize required Pillar 3 accounting, risk, other information, and frequency is stated quarterly, semi-annually and annually.

The main objective of this paper is to analyse the part of the commercial bank website related to Pillar 3 and learn about stakeholders' behaviour. We try to answer the question: Is this "information opportunity" interesting to different kinds of stakeholders?

The structure of this paper is as follows. A summary of the current status of the research is given in the chapter two. In chapter three are methods used and data resources. The fourth and fifth chapters describe the experimental results in detail. The results obtained and conclusions reached are presented in the last chapter.

2. Related Work

The vast majority of literature, and existing studies, deal with the topic of market discipline in the banking industry in the context of mature economies. In addition to the definition presented in this paper, other authors conceptually explain MD in banking in two different ways: direct market discipline and indirect market discipline. Three types of instruments should be used to implement MD [8]: impose more transparency, change the capital structure of banks, use market information.

Various literature sources present the opinion that Pillar 3 should be the "instrument" forcing bank managers to disclose publicly various information. However, many academicians and practitioners believe that the last financial crisis depicted serious shortcomings of Pillar 3. Stephanou [1] studied MD during the crisis, and his opinion is that it is inevitable to rethink MD in banking based upon lessons from the financial crisis. The key issue for him is how the MD concept under different financial system structures and institutional environments can be operationalized. He proposed to look at this issue from four interrelated building block perspectives taking into consideration macroeconomic environment and financial system structure: 1. Information and disclosure, 2. market participants, 3. discipline mechanism, 4. internal governance.

According to [1]: "This would imply significant reforms in Pillar 3 of Basel II, which implicitly assumes that the preconditions already exist for the market to utilize information effectively to exercise discipline... . Unless such reforms are undertaken, Pillar 3 will remain the weakest pillar of the Basel II Framework since it suffers from too many structural limitations to be of much use for MD purposes." Similar a conclusion came from [9], and the other authors e.g. [10]. By developing this distinction between disclosure and transparency, we expect to clarify the role of information transmission when market discipline plays a limited role as happened during the recent crisis.

To fix these limitations additional research is required that would materialize in relevant policies. Stakeholders of the banks, one of key building blocks have been studied from different perspectives. However, they have not been studied much as far as their behaviour in relation to usage of Pillar 3 information during the recent financial crisis. The European Banking Authority (EBA) in its regular survey on Pillar 3 implementation [11] used also questionnaire to improve outcome of Pillar 3 for investors/users. However, they interviewed only rating agencies, credit institutions and analysts –who are not typical representatives of the stakeholders of the commercial bank operating in CEE, owned by a foreign group.

So far there has been no study carried out that would assess the fulfilment of these regulations, especially whether banks' stakeholders, and predominantly clients, really used to a large extent this information to decide to conduct business with a commercial bank during the crisis and still do, and, moreover, whether these regulations help to decrease its vagueness and risk. We analyse the market participants' interest in mandatory disclosure of financial information by means of advanced methods of web log mining.

Web log mining has received extensive attention because of its significant theoretical background and powerful application potential. Many web usage mining approaches and methods were surveyed in [12-14]. The last comprehensive surveys on WUM have been done by Koutri, Avouiris and Daskalaki [15] and Kosala and Blockeel [16].

Facca and Lanzi summarize recent developments in WUM research in their expert studies [17] and [18] where we can observe the progress in this research area. The authors of these papers are of one mind that the web usage mining process can be regarded as a three-phase process, consisting of the data preparation, pattern discovery and pattern analysis phase [19].

In the first phase, web log data are pre-processed in order to identify users, sessions, page views, and clickstreams [20].

Pre-processing refers 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 include erroneous information. According to Cooley, Mobasher and Srivastava [19] it consists of

1. data cleaning (for removing irrelevant references and fields, removing erroneous references, adding missing references due to caching mechanisms, etc.) [21],
2. data integration (for synchronizing data from multiple server logs, integrating registration data, etc.) [22],
3. data transformation (for user-session identification [23-24], path completion[25-26], etc.),
4. and data reduction (for reducing dimensionality) [27-28].

Pattern discovery of web usage mining, which is the outcome of the proposed methodology, is discussed in detail in [29-31]. According to Song et al. [30] behaviour pattern discovering is one of the WLM sub-areas that focuses on finding typical flow models of user actions from recorded events..

Discovering user behaviour patterns is the most frequent application of web log mining. Predominantly used methods are discovering of association and sequence rules, segmentation (cluster analysis, analogy-based methods, etc.) as well as classification (decision rules, decision trees, Bayesian classification, etc.). We model the behaviour of users browsing bank website where information is publicly disclosed under Basel II. The purpose of the analysis is the modelling of website users' behaviour during investigated period.

3. Methods

The BCBS and EU legislation (CRD) only require Pillar 3 information to be disclosed publicly. Banks publish the Pillar 3 information on their websites, which is the best way to make information easily accessible to market participants. We analyse website to learn about stakeholders' behaviour in relation to Pillar 3 and to answer research questions.

When the user visits the website, lot of data is sent to the web server. A common web server retains accesses of a user in the log file and logs basic information about the 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 [32].

Rather than individual page views, we consider complete sequences of views – sessions, which describe the whole of a user's experience on a site so that we can analyse users´ behaviour. Data of outstanding quality requires as much rigorous data gathering as does data preprocessing. We use data from the commercial bank web server which covers a time period from the fourth quarter of the 2008 to the fourth quarter of the 2010. This web server provides bank public information. The selected period was selected by design as it covers website accesses before and during the Recent Financial Crisis.

3.1. Basic data preprocessing

Experiments that analyse user's behaviour usually take a smaller set of data, e.g. two weeks or one month [33-34]. Another approach is to select a few reference weeks [35]. We analyse a quite extensive period (27 months), so we have to alter commonly used methods [36-38] and internal applications for log file preprocessing. We have to rewrite these desktop applications to the server ones so that it can run in batch mode.

The first step in the log file preprocessing is the removal of unnecessary data. These data allow access to graphic files, or style sheets [39]and the accesses from web robots. Different aspects of data preparation can be found in [20] and [27]. Web robots may be defined as autonomous systems that send requests to Web servers across the Internet to request resources. A canonical example of a Web robot is a search engine indexer while a less common example is an RSS feed crawler or a robot designed to collect Web sites for an Internet archive [40]. The difference between the robot and the human can be determined based upon 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, and we can use new methods such as standard deviation of requested page's depth or percentage of consecutive sequential HTTP requests [41]. There are many robots or crawlers that cannot be identified based upon general crawler attributes. In this case, we can use the method navigational patters analysis [42]. After cleaning 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 remains.

Log files also contain information about IP addresses. In cooperation with the bank, we marked IP addresses used in bank local networks. The staff is more familiar with the organization and has a better knowledge of the website structure [43]. We removed these accesses because the majority of accesses are from content creators, web administrators and managers responsible for information disclosure, not from stakeholders or users looking for specific bank information.

3.2. Additional attributes to log file

When we wish to study stakeholders' behaviour, we have to add additional attributes to log file records. Some of them can be added automatically in the application and some require manual assignments. These attributes are as follows:

- Financial quarter (Q_I, Q_{II}, Q_{III}, Q_{IV}) of the year was added to simplify the process of statistical analysis..

- Affiliation of web page to category content. We create 23 categories based upon the content of webpages, e.g. financial reports, annual reports and ethical codex.
- The Pillar 3 relation to the category. Based upon the information on webpages we merge categories into three sets - Pillar 3 Disclosure Requirements, Pillar 3 Related and Other.
- Country or language. The bank website is translated to German and English languages, so we assume that the user reading the non Slovak version of the site is the foreigner. The use of the GeoIP database will not be accurate in the analysis since we are seeking different behaviour between the domestic and foreign user.

3.3. Session identification

Session refers to a series links made by one web-user to get information on a certain topic. At present, there is some arithmetic approach 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 on 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 [44].

For session identification, we use a Session Timeout Threshold (STT) [45]. Other common methods are identification based upon Access Time Threshold and session reconstruction [44]. The aim of session identification is to divide accesses of every user into separate sets. Session can eliminate the side effect (anonymity) of users behind NAT router or proxy server so we can identify users which share one computer, e.g. in the library.

The STT method divides the session into smaller ones when it finds accesses to the web page from the same source with a higher interleave than the set time period – time window. The appropriate size of the time window evolves the quality and quantity of behaviour patterns found. We estimate the value of time windows based upon the duration of the visit - *Length* variable. We calculate the Median, quartile range and non-outlier range of the *Length* variable (Fig. 1a).

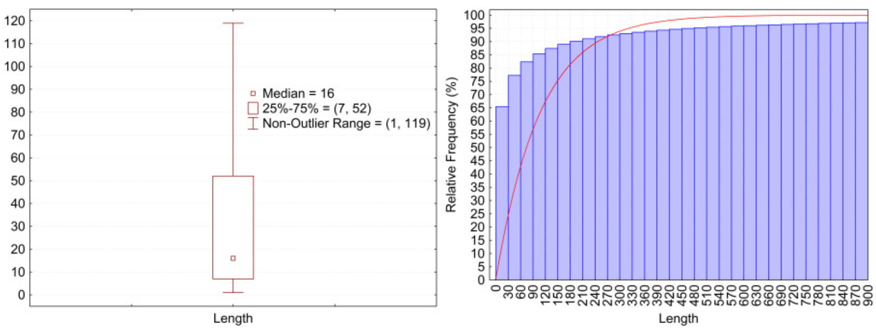

Fig. 1. (a) Box plot: median/quartile range/non-outlier range of the Length (b) Cumulative frequencies of Length variable

The non-outlier range is created between the last value of Q_{III} + *1,5Q* and the last value of Length above Q_I - *1,5Q*. Values outside this interval are considered as outliers.

Table 1. Descriptive characteristics of variable *Length*.

	Mean	Median	Minimum	Maximum	Q_I	Q_{III}	Q_{III}+1,5Q	Q
Length	107.17	16.00	1.00	3600.00	7.00	52.00	119.50	45.00

Our focus is on the upper limit of the non-outlier range. Values that are higher than 119 are considered outliers (Tab. 1). The number 119 is the size of the time window. When the time between accesses to website from the same source is longer, a new session begins. We can also observe that 85% of the values of the Length are smaller than 120 seconds (Fig. 1b).

4. Results

4.1. Analysis of category visit

The crosstabulation (Tab. 2) shows observed interactive frequencies (*Category* x *Language*), as well as relative frequencies expressed as a percentage in rows, columns and total count. The degree of dependence between variables *Category* and *Language* is represented by contingency coefficients, and significance is tested by Pearson chi-square test. The only requirement (validity assumption) of the use of chi-square test is high enough expected frequencies. The condition is violated if the expected frequencies are lower than 5. The validity assumption of chi-square test in these tests is not violated as expected frequencies are high enough (e_{ij} > 2816).

Table 2. A 2-way summary table of observed frequencies: Category x Language

	Category	**Category**	**Category**	∑
Language	**other**	**Pillar3 related**	**Pillar3 disclosure requirements**	
Foreign	195524	27186	7242	229952
Column %	12.79%	32.98%	36.25%	
Row %	85.03%	11.82%	3.15%	
Total %	11.99%	1.67%	0.44%	14.10%
native	1333095	55249	12734	1401078
Column %	87.21%	67.02%	63.75%	
Row %	95.15%	3.94%	0.91%	
Total %	81.73%	3.39%	0.78%	85.90%
∑	1528619	82435	19976	1631030
	93.72%	5.05%	1.22%	100.00%

Contingency coefficients represent the degree of dependency between two nominal variables. The value of coefficient (Category x Language) is approximately 0.15 while 1 means perfect relationship and 0 no relationship. There is a little dependency among access to *About the bank* content and the *Language*, the contingency coefficient is statistically significant (Chi-square = 34517,54; df = 2; p = 0.00). The zero hypothesis is rejected at the 1% significance level, i.e. the number of accesses to particular portal category depends on the language of the content. Results of interaction frequencies - *Category* x *Language* show a low interest for information that is related to Pillar 3. Foreign users are not interested in general bank information, however, they have a higher interest to Pillar 3 related information than domestic stakeholders.

4.2. Visit rate analysis

We deal with the results of association rules analysis in more detail in this chapter. This analysis represents the nonsequential approach to analysed data. We will not analyse sequences, but transactions, i.e. we will not include the time variable into the analysis. The transaction thereafter represents the set of visited web parts of content "About Bank" related to the Pillar 3. Concerning data, we will consider web parts of the commerce bank portal related to the Pillar 3 (Category: Pillar 3, related, Pillar 3 disclosure requirements) which was visited by the stakeholder during one session as a one transaction.

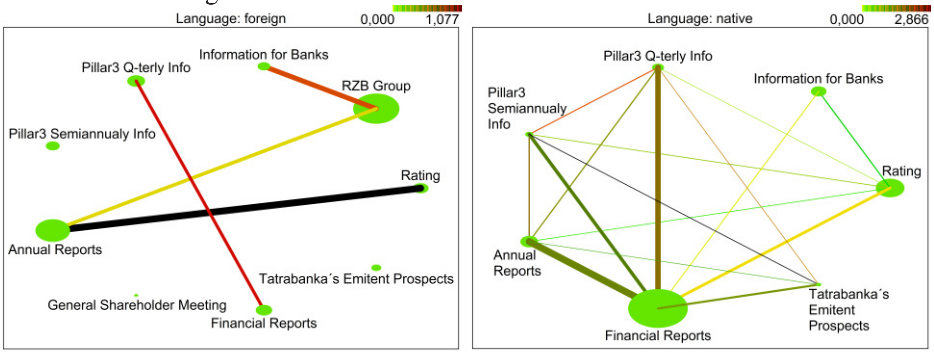

Fig. 2. Web graph: Visualisation of found rules for webpages in the foreign languages (a) and in original/Slovak language (b)

Web graph (Fig. 2a) visualises the association rules found [46-47]. The size of a node represents *support* of this category item, line width represents support of the rule and line brightness stands for the *lift* of the rule. In other words, the *lift* represents the measure of interestingness and offers the most interesting results because it can be interpreted as how many times the categories of problems occur together than in the case if they were statistically insignificant.

Considering the web graph we can see that the most web parts of the portal belong *Group, Annual Reports, Pillar3 Q-terly Info and Financial Reports* (*support* > 10%). The combinations of these web parts in the identified sessions do not exist or exist with very low probability (*support* < 1%),. We can also see that the web parts *Annual Reports ==> Rating* are more frequent together as apart in identified sessions (*lift* = 1.08). We can claim the same for web parts *Pillar3 Q-terly Info ==> Financial Reports* (*lift* = 1.03). We can observe the greatest measure of interestingness (*lift*) in these cases, because the *lift* greater than one means that the selected pairs of web parts are more frequent together than apart in the set of web parts visited by the particular stakeholders. But, we have to be aware that the *lift* does not depend on the rule orientation.

The measure of interestingness (*lift)* of the remaining found rules was less than one.

Considering the web graph (Fig. 2b) we see that web parts *Financial Reports*, *Rating*, *Annual Reports*, *Information for Banks* and *Pillar3 Q-terly Info* (*support* > 10%) belong to the most visited parts of the portal. We found out that the combinations of parts *Annual Reports ==> Financial Reports* and *Pillar3 Q-terly Info ==> Financial Reports* have *support* > 10% at the same time.

The other interesting finding is that *Bank´s Emitent Prospects ==> Pillar3 semiannually Info*, *Pillar3 semiannually Info ==> Pillar3 Q-terly Info*, *Annual Reports ==> Pillar3 semiannually Info* and *Pillar3 Q-terly Info ==> Financial Reports* exist together more frequently than apart (*lift* > 1.6) in the identified sessions.

Another group of rules, *Pillar3 Q-terly Info ==> Bank´s Emitent Prospects*, *Financial Reports ==> Pillar3 semiannually Info*, *Financial Reports ==> Annual Reports*, *Annual Reports ==> Pillar3 Q-terly Info* and *Financial Reports ==> Bank´s Emitent Prospects* attain also the *lift* >1.1. The measure of interestingness (*lift)* of the remaining found rules was less than one.

5. Discussion

The bank studied bank is a typical representative for its peer group in the Slovak market: it is a universal commercial bank. By size is an important financial institution wherein failure could cause serious problems in a local market. The bank has a multiplicity of stakeholders. Among key role players are insured and uninsured depositors, mortgage bond holders, subordinated debtholders and minority shareholders. The main findings of the bank´s stakeholders´ behaviour are presented in the next sections of this chapter.

The stakeholders of the bank exhibited a very low interest in Pillar 3 mandatory required information (1.22%) and information related to Pillar 3 (5.05%) in 2008 and 2009. The highest interest, almost 94%, was on general information on the bank. Possible explanations for this kind of behaviour is follows:

- There was almost no impact of the crisis on the domestic banks as they were in a very good shape as far as capital and liquidity positions. Depositors, as key stakeholders, did not have a serious reason to monitor them.
- Deposit insurance fund coverage, and tough competition among commercial banks operating in the market, might create a sufficient safety net that could be a reason not only for low market discipline, but also for moral hazard in the stakeholders´ behaviour.
- Stakeholders´ structure – numerous minority shareholders have a lack of incentives to control the bank and prefer to rely on control executed by majority shareholder.
- Local regulator (NBS) is demanding, has a very good reputation and stakeholders rely on it.
- Insufficient financial literacy and information „marketing“ that would inform the banks' stakeholders on their role in market discipline - concretely Pillar 3 information.
- Level of sophistication of the presented information (that is more or less standardized by Decree issued by the local regulator) is above of financial literacy and expertise of stakeholders. In this respect also the bank´s complexity of operations have a significant impact on level of sophistication of provided information.

Foreign stakeholders vs. domestic ones presented higher interest in both Pillar 3 disclosure mandatory required information (3.15% vs. 0.91%) and Pillar 3 related information (11.82% vs. 3.94%). It is also evident from the analysis that both groups of stakeholders were more interested in information related to Pillar 3 than Pillar 3 disclosure required information. And above all they were more interested in this type of information provided quarterly than semi-annually. Further analysis discovered that the highest frequency of visiting Pillar 3 quarterly disclosure required information was in the fourth quarter.

The highest interest of foreign stakeholders was at *Group information* and usually in combination with *Information provided for banks*, which indirectly confirms that these stakeholders might be analysts from partners´ banks, rating agencies etc. The second highest frequency (the most significant) had the *Annual reports* category which is tightly related to visit of *Rating* part. Those who visit *Annual Reports also* visit *Group Information* (and vice versa) but less frequently. Foreign stakeholders visit a maximum of two web parts in general.

The highest frequency of domestic stakeholders' visits was the *Financial Reports* part, *Rating*, *Annual Report* and *Information on Banks*. In comparison to foreign stakeholders there are more subsequent visits from one web part than foreign stakeholders do. The maximum is six visits from *Financial Reports* and most often the highest occurrence is between *Financial Reports* and *Annual Reports*, *Financial Reports Pillar 3-Quarterly info*. The outcomes of web information analysis and modelling [48] are patterns of stakeholders' behaviour. There is a difference in the pattern of behaviour between domestic stakeholders and foreign ones. Domestic stakeholders behaviour pattern looks like messy one – stakeholders are fishing for information without knowing what information they are looking for. The foreign stakeholders behaviour pattern is more „focus oriented“ – it looks as if they know what information they are looking for. This different behaviour can be explained by higher literacy of foreign stakeholders (they can be analysts of rating agencies, participating banks etc.).

6. Conclusion

We found out that Pillar 3 disclosure is fully formally implemented in line with BCBS and EBA requirements. However, the lack of the commercial bank´s stakeholders interest on the published information and their pattern of behaviour addressed problems, and raised further research questions. These problems can belong to one of the following two groups: stakeholders have not been able to cope with Pillar 3 disclosed information, or they did not have any incentives. Lack of financial literacy of stakeholders, their information needs and mandatory standardized format of disclosed information by the regulator are the main areas of the first group of problems. The financial safety net in the country, transparency and moral hazard in stakeholder behaviour belong to the second group of problems. In this context it is also important to study the legal regulatory framework that determines content and frequency of presented information in Slovakia: whether it is inevitable to standardize format, and also, whether an annual full set of Pillar 3 disclosure would not be sufficient, instead of quarterly and semi-annually disclosed sets. Change of content and frequency might be beneficial for stakeholders and have a positive impact from time consuming and costs points of view for banks.

References

[1] Stephanou C. 2010. Rethinking market discipline in banking: lessons from the financial crisis 2010. In: *The World Bank*.
[2] BCBS. 2006. International Convergence of Capital Measurement and Capital Standards. In: *Bank for International Settlements*, 2006.
[3] BCBS. 2011. Basel III: A global regulatory framework for more resilient banks and banking systems. In: *Bank for International Settlements*, 2011.
[4] European Parliament and European Council. 2006. Directive 2006/48/EC of 14 June 2006 relating to the taking up and pursuit of the business of credit institutions. In: *Official Journal of the European Union,* 2006, L 177/1 - 197.
[5] European Parliament and European Council. 2006. Directive 2006/49/EC of the European Parliament and of the Council of 14 June 2006 on the capital adequacy of investment firms and credit institutions. In: *Official Journal of the European Union,* 2006, L 177/201.
[6] National Bank of Slovakia. 2010. Decree of the National Bank of Slovakia of 21 September 2010 No. 15/2010 on disclosures by banks and branches of foreign banks.
[7] National Bank of Slovakia. 2011. Decree of the National Bank of Slovakia of 29 November 2011 No. 19/2011 amending Decree of National Bank of Slovakia No 15/2010 on disclosures by banks and branches of foreign banks.
[8] Borio C.E.V. 2004. *Market Discipline Across Countries*. Mit Press.
[9] Deloitte. 2009. *Basel II Pillar 3 - Benchmarking Survey 2009*.
[10] Dewatripont M., Freixas X. 2012. *The crisis aftermath: The new regulatory paradigms*, London: Centre for Economic Policy Research.
[11] European Banking Authority. 2012. *Follow-up review of banks´transparency in their 2011 Pillar 3 reports*.
[12] Srivastava J. et al. 2000. Web usage mining: discovery and applications of usage patterns from Web data. In: *SIGKDD Explor. Newsl*., 2000. 1(2): p. 12-23.
[13] Pierrakos D. et al., 2003. Web Usage Mining as a Tool for Personalization: A Survey. In: *User Modeling and User-Adapted Interaction*. 13(4): p. 311-372.
[14] Tao F., Murtagh F. 2000. Towards Knowledge Discovery from WWW Log Data. In: *Proceedings of the The International Conference on Information Technology: Coding and Computing (ITCC'00) 2000*, IEEE Computer Society. p. 302.
[15] Koutri M., Avouris N., Daskalaki S. 2005. A Survey of Web-Usage Mining: Techniques for Building Web-Based Adaptive Hypermedia Systems. In: *Adaptable and Adaptive Hypermedia Systems 2005*, IGI Global. p. 125-149.
[16] Kosala R., Blockeel H. 2000. Web mining research: a survey. In: *SIGKDD Explor. Newsl.*, 2000. 2(1): p. 1-15.
[17] Facca F., Lanzi P. 2003. Recent Developments in Web Usage Mining Research. In: *Data Warehousing and Knowledge Discovery*. 2003, Springer Berlin Heidelberg. p. 140-150.
[18] Facca F.M., Lanzi P.L. 2005. Mining interesting knowledge from weblogs: a survey. In: *Data Knowl. Eng.*, 2005. 53(3): p. 225-241.
[19] Cooley R., Mobasher B., Srivastava J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns. In: *Knowledge and Information Systems*, 1999. 1(1): p. 5-32.
[20] Chitraa V., Davamani A. 2010. *A Survey on Preprocessing Methods for Web Usage Data*.
[21] Aye T.T. 2011. Web log cleaning for mining of web usage patterns. In: *Computer Research and Development (ICCRD)*, 2011 3rd International Conference on Computer Research and Development.

[22] 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 Data Engineering Workshops.
[23] Chen Z., Fu A.C., Tong F.-H. 2002. Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs. In: *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg. p. 290-296.
[24] Kapusta J., Munk M., Drlik M. 2012. Cut-off time calculation for user session identification by reference length. In: *Application of Information and Communication Technologies (AICT)*, 2012 6th International Conference on.
[25] Zhang C., Zhuang L. 2008. New Path Filling Method on Data Preprocessing in Web Mining. In: *Computer and Information Science*, 2008. 1(3): p. 112-115.
[26] Yan L., Boqin F., Qinjiao M. 2008. Research on Path Completion Technique in Web Usage Mining. In: *Computer Science and Computational Technology*, 2008. ISCSCT '08.
[27] Liu B., Mobasher B., Nasraoui O. 2011. Web Usage Mining. In: *Web Data Mining* 2011, Springer Berlin Heidelberg. p. 527-603.
[28] 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: *From Web to Social Web: Discovering and Deploying User and Content Profiles*. Springer Berlin Heidelberg. p. 82-101.
[29] Shun-Hua T., Miao C., Guo-Hai Y. 2010. User behavior mining on large scale web log data. In: *Apperceiving Computing and Intelligence Analysis (ICACIA)*
[30] Jinliang S., et al. 2007. The Data Preprocessing of Behavior Pattern Discovering in Collaboration Environment. In: *Web Intelligence and Intelligent Agent Technology Workshops*.
[31] Nina S.P., et al. 2009. Pattern Discovery of Web Usage Mining. In: *Computer Technology and Development, ICCTD '09*.
[32] Thomas P. 2012. Explaining difficulty navigating a website using page view data. In: *Proceedings of the Seventeenth Australasian Document Computing Symposium 2012*, ACM: Dunedin, New Zealand. p. 31-38.
[33] Liu H., Kešelj V. 2007. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. In: *Data & Knowledge Engineering*, 2007. 61(2): p. 304-330.
[34] Stevanovic D., An A., Vlajic N. 2012. Feature evaluation for web crawler detection with data mining techniques. In: *Expert Systems with Applications*, 2012. 39(10): p. 8707-8717.
[35] Xing D., Shen J. 2004. Efficient data mining for web navigation patterns. In: *Information and Software Technology*, 2004. 46(1): p. 55-63.
[36] Munk M., Kapusta J., Svec P. 2010. Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor, In: *ICCS 2010 - International Conference on Computational Science*, Proceedings 2010, Elsevier Science Bv: Amsterdam. p. 2267-2274.
[37] Munk M., Kapusta J., Svec P. 2009. Data preprocessing dependency for web usage mining based on sequence rule analysis. In: *IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2009*. Algarve, Portugal.
[38] Munk M, Drlik M. 2011. Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System. In: *Procedia Computer Science*, 2011. 4(0): p. 1640-1649.
[39] Ortega J.L., Aguillo I. 2010. Differences between web sessions according to the origin of their visits. In: *Journal of Informetrics*, 2010. 4(3): p. 331-337.
[40] Doran D., Gokhale S.S. 2011. Web robot detection techniques: overview and limitations. In: *Data Mining and Knowledge Discovery*., 2011. 22(1-2): p. 183-210.
[41] Stevanovic D., An A., Vlajic N. 2011. Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers, In: *Foundations of Intelligent Systems*. 2011, Springer Berlin Heidelberg. p. 483-489.
[42] Tan P.-N., Kumar V. 2002. Discovery of Web Robot Sessions Based on their Navigational Patterns. In: *Data Mining and Knowledge Discovery*, 2002. 6(1): p. 9-35.
[43] Thomas P., Paris C. 2010. Interaction differences in web search and browse logs. In: *Proceedings of the 15th Australasian Document Computing Symposium 2010*. Melbourne, Australia.
[44] Fang Y., Huang Z. 2010. An Improved Algorithm for Session Identification on Web Log. In: *Web Information Systems and Mining 2010*, Springer Berlin Heidelberg. p. 53-60.
[45] Munk M., Drlik M. 2011. Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining. Dijon. p. 60-74.
[46] Klocokova D. 2011. Integration of heuristics elements in the web-based learning environment: Experimental evaluation and usage analysis. In: *Procedia Social and Behavioral Sciences*, 2011. 15: p. 1010-1014.
[47] Capay M., Balogh Z., Boledovicova M., Mesarosova, M. 2011. Interpretation of Questionnaire Survey Results in Comparison with Usage Analysis in E-Learning System for Healthcare. In: *Communications in Computer and Information Science*, 2011. 167(2): p. 504-516.
[48] Skorpil, V., Stastny, J. 2008. Comparison of Learning Algorithms. In: *24th Biennial Symposium on Communications*, Kingston, Canada, 2008. p. 231-234.