ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LX 29 Number 2, 2012

VERIFICATION OF THE FULFILMENT OF THE PURPOSES OF BASEL II, PILLAR 3 THROUGH APPLICATION OF THE WEB LOG MINING METHODS

M. Munk, A. Pilková, M. Drlík, J. Kapusta, P. Švec

Received: November 30, 2011

Abstract

MUNK, M., PILKOVÁ, A., DRLÍK, M., KAPUSTA, J., ŠVEC, P.: *Verification of the fulfilment of the purposes of Basel II, Pillar 3 through application of the web log mining methods.* Acta univ. agric. et silvic. Mendel. Brun., 2012, LX, No. 2, pp. 217–222

The objective of the paper is the verification of the fulfilment of the purposes of Basel II, Pillar 3 – market discipline during the recent financial crisis. The objective of the paper is to describe the current state of the working out of the project that is focused on the analysis of the market participants' interest in mandatory disclosure of financial information by a commercial bank by means of advanced methods of web log mining. The output of the realized project will be the verification of the assumptions related to the purposes of Basel III by means of the web mining methods, the recommendations for possible reduction of mandatory disclosure of information under Basel II and III, the proposal of the methodology for data preparation for web log mining in this application domain and the generalised procedure for users' behaviour modelling dependent on time. The schedule of the project has been divided into three phases. The paper deals with its first phase that is focusing on the data pre-processing, analysis and evaluation of the required information under Basel II, Pillar 3 since 2008 and its disclosure into the web site of a commercial bank. The authors introduce the methodologies for data preparation and known heuristic methods for path completion into web log files with respect to the particularity of investigated application domain. They propose scientific methods for modelling users' behaviour of the webpages related to Pillar 3 with respect to time.

web log mining, user behaviour model, multinomial logit model, Basel Accords

Basel II is the second of the Basel Accords, which are recommendations on banking laws and regulations issued by the Basel Committee on Banking Supervision. The purpose of Basel II is to create an international standard that banking regulators can use when creating regulations about how much capital banks need to put aside to guard against the types of financial and operational risks banks face while maintaining sufficient consistency so that this does not become a source of competitive inequality amongst internationally active banks (Group of Governors and Heads of Supervision announces higher global minimum capital standards, 2010).

Basel II introduces the concept of three pillars:

- Pillar 1: Minimum Capital Requirements (addressing risk),
- Pillar 2: Supervisory Review,
- Pillar 3: Market Discipline.

The third pillar – Market Discipline – was established as one of the complementary elements of The New Basel Capital Accord in 2004 by Basel Committee on Banking Supervision and later adopted in binding European legislation (Directive 2006/48, 2006).

In accordance with this document the information are mandatory disclosed by banks quarterly and on semi-annual basis. According to the information released the stakeholders can consider banks' risk exposure, management and capital adequacy. The objective of this regulation was to provide possibility to consider relevance of banks' risk management for stakeholders and consequently to put banks with high risk appetite under pressure through stakeholders' deposits allocation with requirement of higher spread.

Since the recent financial crisis revealed shortcomings in this field, Basel Committee on Banking Supervision issued document known as Basel III (Basel III, 2010).

Basel III is a new global regulatory standard on bank capital adequacy and liquidity agreed upon by the members of the Basel Committee on Banking Supervision. The third of the Basel Accords was developed in a response to the deficiencies in financial regulation revealed by the global financial crisis. Basel III strengthens bank capital requirements and introduces new regulatory requirements on bank liquidity and bank leverage (Basel III, 2010). Inevitable part of this document is enlargement of the information scale mandatory disclosed by banks.

So far there has been no study carried out that would assess the fulfilment of this regulations, especially whether banks' stakeholders and predominantly clients really used to a larger 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.

There can be no doubt of the fact that Basel II as well as Pillar 3 meant the increase in operating and capital costs for the banks. This begs a question: "Is the regulation adequate or does it only causes the increase in costs compensation of which urges banks to take more risk that may contribute to crisis state?" It is really demanding to find the answer and it requires extensive data analysis of websites of particular commercial banks.

Therefore, the main objective of the paper is to present proposed methodology of analysis such kind of website with the emphasis placed on data preparation and to describe the model of website visitors' behaviour of particular commercial banks.

Web mining

The aim of this paper lies in the verification of the fulfilment of the purposes present in Basel II, Pillar 3 – Market Discipline during the recent financial crisis on the basis using advanced web mining methods. Web mining closely related to the research field that is known as knowledge discovery in databases (KDD). Web mining can be defined as an extraction of interesting and potentially useful knowledge and information from activities referring to World Wide Web (Liu, 2007).

Sometimes, for the application of web mining methods, it is sufficient only to slightly adjust the existing procedures from the scope of KDD, otherwise it is necessary to change steps of data advance preparation and transformation more radically. Web mining can be divided into three domains (Zaiane and Han, 1998):

- web content mining,
- web structure mining,
- web usage mining.

Web usage mining is focused on the analysis of behaviour of users while surfing the net. The most frequent sources of data are the ones automatically stored in the web log files, so web usage mining is often identified with web log mining.

RESEARCH METHODOLOGY

Data obtained from the log files of the bank website are raw data that have to be pre-processed before it can be further analysed using web log mining methods. The proposed methodology will consider the objectives of the research and will be realized in four steps:

1. Data acquisition – defining the observed variables into the log file from the point of view of obtaining the necessary data (IP address, date and time of access, URL address, etc.).
2. Data preparation that is described in the next chapter in more detail.
3. Data analysis – web users' behaviour modelling.
4. Interpretation of results and summing up of recommendations for website administration and further research.

Data preparation

High quality of data is essential for the good data analysis. Data preparation is probably the longest and the most time-consuming phase in the process of web log mining because of incompleteness of accessible data as well as irrelevant information present in the collected data.

When a user visits a website a lot of information is sent to a server. Most of the web servers automatically store access records, i.e. logs. The log file of the web server is the source of anonymous data about the user. But, these anonymous data present problem by unambiguous identification of the particular user.

If there is a request to know whether a user visits particular website, browses through publicly disclosed information and how much time the user spends on the website, it is necessary to know how to detach his/her activities from activities of other users while preserving his anonymity.

The automatically saved data in log file are data source that represents input to further analysis. Log file, in its standard structure called Common Log File marks each transaction performed by the browser in accessing the web. Each row presents notice about IP address, the time and the date of the

visit, the requested and referring object (W3C 1995). It is sometimes better to use its extended version that also records the version of the browser, User-Agent.

Different aspects of data preparation can be found, for example, in work by Chitraa and Davamani (Chitraa and Davamani, 2010) and Bing (Bing, 2006).

The data preparation for the needs of proposed research of web pages containing information about the third pillar of Basel II consists of four steps. In short, the following adjustments (corrections) are made:

- Data cleaning from the crawlers of search services accessed to the portal (Lourenço and Belo, 2006).
- Identification of visitors based on the various internet browsers (Cooley, Mobasher and Srivastava, 1999).
- Identification of sessions, where the session may be defined as a sequence of the steps, that lead to completing the concrete task or as a sequence of the steps, that lead to meeting the concrete target (Chen, Park and Yu, 1996). The simplest method is based on considering the series of clicks in a defined period of time, for example 15 minutes (Spiliopoulou and Faulstich, 1999). The comprehensive research on the impact of different session time thresholds can be found in other papers of the authors (Munk and Drlík, 2011).
- The reconstruction of activities of a web visitor. Taucher and Greenberg (Taucher and Greenberg, 1997) proved that more than 50% of accesses to web pages are via backward path. Here comes the problem with the cache of the browser. By the backward path, a request to web server is not ran, thus there does not exist a record in the log file. The solution of this problem is path reconstruction. The path reconstruction adds these missing records into the log file (Cooley, Mobasher and Srivastava, 1999). As a result, reconstruction of the activity of every website user is demanding procedure from perspective of theory, time and technical realisation.

The methodology of the data preparation used in this paper is closely related to the authors' previous experiments in different application domains, in which authors assessed the relevance of respective steps of data preparation for a further detailed analysis (Munk, Kapusta and Švec, 2010; Munk *et al.* 2010; Munk and Drlík, 2011). The resulting refined log file is an appropriate input for a detailed analysis of the bank website visitors' behaviour.

Discovering user's behaviour patterns

Discovering of website users' behaviour pattern belongs to the most frequent applications 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.) (Domenech and Lorenzo, 2007).

In this paper the authors 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 of time. Therefore, two methods are used – a sequence rule analysis and a multinomial logit model.

The sequence rule analysis belongs to the standard web log mining methods. Therefore, the authors have decided to pay more attention in this paper to the multinomial logit model, the usage of which is untypical for this application domain and can be used for the users' behaviour modelling depending on time.

Multinomial logit model

From the perspective of the web log mining the collected data can be seen as time data. None of mentioned web log mining methods model users' behaviour depending on time and, moreover, according to information gained it has not been sufficiently described in scientific literature yet. The multinomial logit model can be described as follows.

Denote by π_{ij} the probability of choosing the web portal category j by a visitor in the hour i, where $j = 1, 2, \ldots, J$. Since $\sum_{j=1}^{J} \pi_{ij} = 1$, there are $J - 1$ parameters.

Let Y_{ij} be the number of accesses in the hour i to category j with observed value y_{ij} Then $n = \Sigma_j y_{ij}$ is the number of accesses in the hour i. The probability distribution of Y_{ij}, in the case n_i is given, is multinomial,

$$P[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iJ} = y_{iJ},] = \frac{n_i!}{y_{i1}!\, y_{i2}! \ldots y_{iJ}!} \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \ldots \pi_{iJ}^{y_{iJ}}. \quad (1)$$

The probabilities π_{ij} of choosing category j with respect to the hour i we get from multinomial logits which we will model. The multinomial logits are logarithms

$$\ln \frac{\pi_{ij}}{\pi_{iJ}},\ j = 1, 2, \ldots, J-1,\ i \in \{0, 1, \ldots, 23\},$$

where π_{iJ} is the probability of last (reference) category. We assume the following model

$$\eta_{ij} = \ln \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \mathbf{x}_i' \beta_j, \quad (2)$$

where $\mathbf{x}$' is a line vector, α_j is a constant and β_j is a vector of regression coefficients, for j = 1, 2, …, $J - 1$. There are $J - 1$ equations which contrast each of categories 1, 2, …, $J - 1$ with the category J. The probabilities π_{ij} we obtain from formulas

$$\pi_{iJ} = \frac{1}{1 + \sum_{k=1}^{J-1} e^{\eta_{ik}}},\quad \pi_{ij} = e^{\eta_{ij}} \pi_{iJ},\ j = 1, 2, \ldots, J-1. \quad (3)$$

Maximum likelihood estimation of the parameters of the model (2) proceeds by maximization of the multinomial likelihood (1) with the probabilities π_{ij} viewed as functions of α_j and β_j. The estimation of the parameters can be done for individual data in a statistical system. Then $n_i = \Sigma_j y_{ij} = 1$ for where n is the number of all accesses.

It is necessary to emphasize that this method, contrary to standard used methods of web log mining, enables to model probabilities of accesses depending on time (Munk, Vrábelová and Kapusta, 2011; Munk, Drlík and Vrábelová, 2011).

DISCUSSION AND CONCLUSIONS

The authors present modelling possibilities of the distribution of categorical variable with the support of a multinomial logit model. The paper introduces only the basic starting points and objectives of the project proposal and describe authors' own model of user's behaviour in time in more detail due to the overall complexity of the problem.

Even though, the contribution of the paper can be described in more details from a perspective of the application domain as well as from a perspective of web log mining. Contribution to application domain can be summarized to these points:

- The analysis of the relevant problem, i.e. research into fulfilment of purposes of Basel II regulation, Pillar 3 – Market Discipline. So far no systematic research into this topic has been conducted neither in Slovakia or abroad. Nevertheless, one of three new European Supervisory Authorities – the European Banking Authority (EBA) is interested in stakeholders' opinion on efficiency of bank regulation. From this perspective the results would be beneficial not only for the theory, but also for practical purposes of European regulation.
- The analysis of the relevant problem by means of genuine and, for this field of study, untraditional methods of web log mining.
- Recommendations for a change of extent and structure of mandatory disclosure of information on the bank's web site.

The research will also contribute to the web log mining research area in the term of identification of steps necessary for reliable data preparation for its later processing by web log mining methods, generalisation of procedure for modelling users' behaviour depending on time, verification and assessment of usage of multinomial logit model for modelling probabilities of accesses depending on time and huge data where high website traffic is expected.

Based on the found users' behaviour patterns, it is possible to modify and improve web page of the bank institution. At the same time, according to the described objectives the change of content, dissemination or accessibility of information on Basel II, Pillar 3 on the website can be managed more precisely and effectively.

SUMMARY

The objective of the paper is the verification of the fulfilment of the purposes of Basel II, Pillar 3 – market discipline during the recent financial crisis. The objective of the paper is to describe the current state of the working out of the project that is focused on the analysis of the market participants' interest in mandatory disclosure of financial information by a commercial bank by means of advanced methods of web log mining. The output of the realized project will be the verification of the assumptions related to the purposes of Basel III by means of the web mining methods, the recommendations for possible reduction of mandatory disclosure of information under Basel II and III, the proposal of the methodology for data preparation for web log mining in this application domain and the generalised procedure for users' behaviour modelling dependent on time. The schedule of the project has been divided into three phases. The paper deals with its first phase that is focusing on the data pre-processing, analysis and evaluation of the required information under Basel II, Pillar 3 since 2008 and its disclosure into the web site of a commercial bank. The authors introduce the methodologies for data preparation and known heuristic methods for path completion into web log files with respect to the particularity of investigated application domain. They propose scientific methods for modelling users' behaviour of the webpages related to Pillar 3 with respect to time.

REFERENCES

Basel III. A Global Regulatory Framework for More Resilient Banks and Banking Systems, 2010: Basel Committee for Banking Supervision [cit. 2011-11-02]. Cited from http://www.bis.org/publ/bcbs189.htm.

BERENDT, B. and SPILIOPOULOU, M., 2000: Analysis of Navigation Behaviour in Web Sites Integrating Multiple Information Systems. *The VLDB Journal*, 2000, Vol. 9, No. 1, pp. 56–75. ISSN 1066-8888.

BING, L., 2006: *Web Data Mining. Exploring Hyperlinks, Contents and Usage Data*. Springer, ISBN 13-978-3-540-37881-5.

COOLEY, R., MOBASHER, B. and SRIVASTAVA J., 1999: Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information System*, 1999, Springer-Verlag, Vol. 1, ISSN 0219-1377.

Directive 2006/48/EC, 2006. Directive 2006/48/EC of the European Parliament and of the Council of 14 June 2006. [cit. 2011-11-02]. Cited from: *http://ec.europa.eu/internal_market/bank/regcapital/index_en.htm.*

DOMENECH, J. M. and LORENZO, J., 2007: A Tool for Web Usage Mining. In: *8th International Conference on Intelligent Data Engineering and Automated Learning*, IDEAL.

Group of Governors and Heads of Supervision announces higher global minimum capital standards. 2010: Basel Committee for Banking Supervision. [cit. 2011-11-02]. Cited from http://www.bis.org/press/p100912.htm. Press release.

CHEN, M., PARK, J. S. and YU, P. S., 1996: Data Mining for Path Traversal Patterns in a Web Environment. In: *ICDCS*, pp. 385–392. ISBN 0-8186-7398-2.

CHITRAA, V. and DAVAMANI, A. S., 2010: A Survey on Preprocessing Methods for Web Usage Data. I*nternational Journal of Computer Science and Information Security*, Vol. 7, Issue 3.

LIU, B. 2007: *Web data mining: Exploring hyperlinks, contents and usage data*. Springer, ISBN 978-3-540-37881-5.

LOURENÇO, A. G. and BELO, O. O., 2006: Catching Web Crawlers in the Act. In: *Proceedings of the 6th international Conference on Web Engineering, ICWE '06*, Vol. 263, ACM, New York, NY, pp. 265–272. ISBN 1-59593-352-2.

MUNK, M. and DRLÍK, M., 2011: Influence of Different Session Timeouts Thresholds on Results of Sequence Rule Analysis in Educational Data Mining. In: *Digital Information and Communication Technology and Its Applications*. Springer Berlin Heidelberg, 2011, Vol. 166, p. 60–74.

MUNK, M. and DRLÍK, M., 2011: Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System. In: *International Conference on Computational Science* 2011, ICCS 2011, Procedia Computer Science, Elsevier.

MUNK, M., DRLÍK, M. and VRÁBELOVÁ, M., 2011: Probability Modelling of Accesses to the Course Activities in the Web-based Educational System. In: *International Conference on Computational Science and its Applications 2011*, ICCSA 2011, Lecture Notes in Computer Science, Springer, 2011.

MUNK, M., KAPUSTA, J. and ŠVEC, P., 2010: Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor. *Procedia Computer Science*, 2010, Vol. 1, No. 1. pp. 2267–2274.

MUNK, M., KAPUSTA, J., ŠVEC, P. and TURČÁNI, M., 2010: Data Advance Preparation Factors Affecting Results of Sequence Rule Analysis in Web Log Mining. *E a M: Ekonomie a Management*, 2010, Vol. 13, No. 4, p. 143–160.

MUNK, M., VRÁBELOVÁ, M. and KAPUSTA, J., 2010: Probability Modelling of Accesses to the Web Parts of Portal. WCIT 2010, *Procedia Computer Science*, Elsevier, Vol. 3, 2011, 677–683.

ROMERO, C., VENTURA, S., ZAFRA, A. and de BRA, P., 2009: Applying Web Usage Mining for Personalizing Hyperlinks in Web-based Adaptive Educational Systems. *Computers & Education*.

SPILIOPOULOU, M. and FAULSTICH, L. C., 1999: WUM: A Tool for Web Utilization Analysis. In: *Extended version of Proc. EDBT Workshop WebDB'98*, 1999, Springer Verlag, pp. 184–203.

TAUCHER, L. and GREENBERG, S., 1997: Revisitation Patterns in World Wide Web Navigation. In: *Proc. of Int. Conf. CHI'97*, 1997, Atlanta.

W3C, 1995: *Configuration File of W3C httpd*. [cit. 2011-11-02]. Cited from: http://www.w3.org/Daemon/User/Config/Logging.html.

ZAIANE, O. and HAN, J., 1998: WebML: Querying the World-Wide Web For Resources and Knowledge. In: *Workshop on Web Information and Data Management*. 1998. p. 9–12.

Address

Mgr. Martin Drlík, PhD., Katedra informatiky, Fakulta prírodných vied, Univerzita Konštantína Filozofa v Nitre, Tr. A. Hlinku 1, 949 74 Nitra, Slovenská republika, e-mail: mdrlik@ukf.sk