Method Article

Methodology of stakeholders’ behaviour modelling based on time

Michal Munk\textsuperscript{a}, Anna Pilkova\textsuperscript{b}, Lubomir Benko\textsuperscript{a,∗}, Petra Blazekova\textsuperscript{b}, Peter Svec\textsuperscript{a}

\textsuperscript{a}Constantine the Philosopher University in Nitra, Slovakia
\textsuperscript{b}Comenius University in Bratislava, Slovakia

A B S T R A C T

The methods presented in this article were created to model and describe the behaviour of the web users of a bank institution web portal. The source dataset is represented by a log file of the commercial bank web server. The analysis is oriented on examining the behaviour of visitors over an extended period (2009-2012). The years 2009-2010 represent the years of the financial crisis, and the years 2011-2012 represent the years after the financial crisis. The following method describes the sequence of steps necessary to pre-process the raw log file and model the web user behaviour using the multinomial logit model. The introduced methods can be used also for other domains in the case of appropriate data preparation.

- Data preparation- data cleaning, user/session identification, path completion, variables determination;
- Data analysis- model definition, parameters estimation, logits estimation, probabilities estimation;
- Results evaluation- comparison of empirical and theoretical values in terms of counts, probabilities and logits.

© 2021 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

A R T I C L E  I N F O

Method name: Methodology of Stakeholders’ Behaviour Modelling Based on Time
Keywords: Data pre-processing, Web usage mining, Multinomial logit model
Article history: Received 6 July 2021; Accepted 31 October 2021; Available online 2 November 2021

Specification table

| Subject Area | Computer Science |
|--------------|------------------|
| More specific subject area | Web Usage Mining |
| Method name | Methodology of Stakeholders’ Behaviour Modelling Based on Time |
| Name and reference of original method | Web usage data pre-processing and analysis [1,2] |
| Resource availability | The pre-processed log file is located in Data in Brief [3]. |

DOI of original article: 10.1016/j.eswa.2021.115503

* Corresponding author.
E-mail address: lbenko@ukf.sk (L. Benko).

https://doi.org/10.1016/j.mex.2021.101570
2215-0161/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
Methods described in this article were created to analyse and model the behaviour of the visitors of a web portal – bank web portal for a research article [4]. The data source was log files obtained from the webservers and it contained the visitors’ accesses to the web parts of the web portal. A detailed description of the log files is located in Data in Brief [3]. Data related to Pillar 3 were gathered from bank webservice log files [5]. The research methodology was inspired by [6–8]. The model of bank visitors’ behaviour was created based on a multinomial logit model [9]. The model was used to determine categorical dependent variable is a variable category that represents a group of web parts that deal with a similar issue. The variable contains these categories: Pricing List, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements and We support. The analysis is orientated on examining the behaviour of visitors over an extended period of time (2009-2012). The years 2009-2010 represent the years of financial crisis (variable crisis=1). On the other hand, the years 2011-2012 represent the years after the financial crisis (variable crisis=0). The time independent variable was chosen as the variable week that was created from the date of the access of the visitor. The variable was created based on the standard ISO 8601 and the variable acquires the values 0-53. If the week number equals 0, it means that the given date belongs to the preceding year. The applied methodology is based on [10] and is as follows:

1. Obtaining log files from multiple servers.
2. Data preparation involving the following multiple tasks:
   a. Data cleaning – removing the unnecessary data from the log files (requests for pictures, styles and so on; and also accesses of robots of search engines) which leads to raw data of only accesses to the web portal.
   b. User/session identification – the visitors were identified based on the variables IP address and user agent; and sessions were identified based on the Reference Length method.
   c. Path completion – used to complete the records of the users’ path that the user followed using the Back button in the web browser (these visited pages are not recorded in the log file since they have already been stored on the client side under the previous steps).
   d. Variables determination – the log file contains the variables in a typical Extended Log Format (ELF), so a transformation and variable definition are needed for the user behaviour analysis of the examined web portal. A dependent variable category is created, and it represents the web parts of the web portal. In case the web parts have low traffic, it is appropriate to create wider categories based on their relevance to the content [9]. The variable category will in the case of the examined web portal of the bank institution contain the following web categories: Pricing List, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements, and We support. It is also necessary to identify independent variables—predictors that represent the time variables created from the timestamp of the access to the web category. In case of the weeks of the year, it is the variable week that was created based on ISO 8601 and will have values of 0-53. The variable will be 0 in case it is a week that begins in the previous year. The next predictor will be the dummy variable crisis that identifies the period of years during a financial crisis and after a financial crisis. Next a dummy variable internal was created for the identification of accesses from inside and outside of the organization’s network. In this way the behaviour of users accessing from the inside/external access (internal/external access) of the organizations’ network (the variable was created based on the sets of IP addresses) can be analysed.
3. Data analysis follows based on the presumption that the examined data consists of individual accesses to web portal parts:
   a. Model definition - probability distribution of accesses \(Y_{ij}\) in time \(i\) for the category \(j\) with observations \(y_{ij}\), if the count of accesses is given \(n_i = \sum_j y_{ij}\) in time \(i\) is multinomial \(P[Y_{11} = y_{11}, Y_{12} = y_{12}, \ldots, Y_{y} = y_{y}] = \frac{n!_{y1}}{y_{11}! \cdots y_{1y}!} \pi_{y1}^{y_{11}} \pi_{y2}^{y_{12}} \cdots y_{yj}^{y_{1y}}\). Since \(\sum_{j} \pi_{ij} = 1\) it is necessary to estimate \(J - 1\) of unknown probabilities. The estimates are calculated using the Maximum Likelihood method. In the logarithmic function of likelihood (without constants) \(\sum_{i} \sum_{j=1}^{y} y_{ij} \ln \pi_{ij}\) (1) is denoted a logit transformation \(\eta_{ij} = \ln \frac{\pi_{ij}}{\pi_{ij}}\), where the last category
is chosen as the reference category \( \eta_{ij} = 0 \), and it is assumed that the logits \( \eta_{ij} \) are linear functions of the independent variables \( \eta_{ij} = \alpha_j + \mathbf{x}_i^T \mathbf{\beta}_j \). Using inverse transformation, it is denoted \( \pi_{ij} = \frac{1}{1 + \sum_{j=1}^{J} \bar{e}^{\eta_{ij}}} \), \( \pi_{ij} = \frac{1}{1 + \sum_{j=1}^{J} \bar{e}^{\eta_{ij}}} \) for \( j = 1, 2, \ldots, J - 1 \), respectively \( \pi_{ij} = \frac{1}{1 + \sum_{j=1}^{J} \bar{e}^{\eta_{ij}}} \). \( \pi_{ij} = \frac{1}{1 + \sum_{j=1}^{J} \bar{e}^{\eta_{ij}}} \), \( j = 1, 2, \ldots, J - 1 \) (2). The logarithm function is a likelihood function with unknown parameters \( \alpha_j \) and \( \mathbf{\beta}_j \) (\( j = 1, 2, \ldots, J - 1 \)) after substituting such expressed \( \pi_{ij} \) (2) into (1). After determining the model, it is necessary to identify the type of dependence for determining the degree of polynomial and the selection of predictors, including dummy variables.

b. The estimation of the models' parameters \( \alpha_j, \mathbf{\beta}_j \) by maximizing the logarithm of the multinomial likelihood function. The STATISTICA Generalized Linear/Nonlinear Models was used to estimate the parameters of individual values. The significance of parameters was tested using the Wald test \( H_0: \alpha_j = 0, H_0: \beta_{kj} = 0 \), where \( k \) is the number of predictors. The estimated parameters are used to calculate estimates of logits and from logits can be calculated probabilities of selection of specific categories at a given time.

c. The estimation of logits \( \hat{\eta}_{ij} \) for all values of independent variables \( \hat{\eta}_{ij} = a_j + \mathbf{x}_i^T \mathbf{b}_j, j = 1, 2, \ldots, J - 1 \).

d. Probability estimation of accesses \( \pi_{ij} \) in time \( i \) for reference web category \( j \) \( \hat{\pi}_{ij} = \frac{1}{1 + \sum_{j=1}^{J} \bar{e}^{\hat{\eta}_{ij}}} \).

e. Probability estimation of accesses \( \pi_{ij} \) in time \( i \) for web category \( \hat{\pi}_{ij} = e^{\hat{\eta}_{ij}} \), \( j = 1, 2, \ldots, J - 1 \).

f. Visualization of the probabilities of web category \( j \) in time \( i \), where \( j = 1, 2, \ldots, J \).

4. Results evaluation:

Based on the assumption that the expected counts \( \hat{y}_{ij} = n_i \hat{\pi}_{ij} \) are big enough (they are not zero and no more than 20% from \( \hat{y}_{ij} \) is less than 5) to compare the actual model with the saturated model that is used to predict the probabilities independently for \( i = 0, 1, \ldots, 53 \), then the statistics \( G^2 \) (deviance, likelihood ratio) can be used

\[
G^2 = LR(\hat{\pi}) = 2(L(p) - L(\hat{\pi})) = 2 \sum_{i=0}^{53} \sum_{j=1}^{J} y_{ij} \ln \left( \frac{p_{ij}}{\hat{\pi}_{ij}} \right) = 2 \sum_{i=0}^{53} \sum_{j=1}^{J} y_{ij} \ln \left( \frac{y_{ij}}{\hat{y}_{ij}} \right).
\]

After the last edit it is following:

\[
G^2 = 2 \sum_{i=0}^{53} \sum_{j=1}^{J} y_{ij} \ln \left( \frac{y_{ij}}{\hat{y}_{ij}} \right).
\]

The hypothesis \( H_0: \pi_{ij} = \hat{\pi}_{ij} \) can be tested using the LR test. LR test makes it possible to compare the estimates \( \hat{y}_{ij} \) with \( y_{ij} \). The saturated model has \( 54(J - 1) \) free parameters and the current model \( k(j - 1) \), then the degrees of freedom \( df \) are equal \( (54 - k)(j - 1) \). Statistics \( G^2 \) has approximately \( \chi^2(df) \) distribution. Pearson statistics can be used to compare the estimates \( \hat{y}_{ij} \) with \( y_{ij} \) either:

\[
\chi^2 = \sum_{i=0}^{53} \sum_{j=1}^{J} r_{ij}^2.
\]

Where \( r_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{y}_{ij}}} \) is Pearson residue that has also the \( \chi^2(df) \) distribution.

In the given application field is often the condition of using the LR test/ Pearson statistics violated. Usually, the examined variable has a considerable number of levels that are web parts of the portal or system (pages, content categories, activities, etc.). This results in the violation of using the LR test/Pearson statistics- the expected counts are not large enough. For this reason are used alternative methods to evaluate the model [9,11] - visualization of empirical and theoretical counts differences, extreme identification, comparison of the distribution of empirical relative counts of accesses and estimated probabilities of the examined web part \( j \) in time \( i \), and empirical and theoretical logit visualization of each web part, except the reference web part.
The created model was evaluated based on the following steps:

a. Empirical counts determination \( y_{ij} \).

b. Theoretical counts estimation \( \hat{y}_{ij} = \hat{\pi}_{ij} \sum_j y_{ij} \).

c. Visualization of differences in the empirical and theoretical counts of accesses \( d_{ij} = y_{ij} - \hat{y}_{ij} \).

d. Extreme values identification \( d_{ij} \), where \( d_{ij} > \hat{d}_j + 2s_j \) represents an underestimated prediction and \( d_{ij} < \hat{d}_j - 2s_j \) represents overestimated prediction where \( s_j \) is standard deviation and \( \hat{d}_j \) is the mean of differences of the category \( j \).

e. Calculation of relative empirical counts of accesses \( p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}} \).

f. Comparison of the distribution of the relative empirical counts of accesses with the estimated probabilities of the selected web category \( j \) in time \( i \). To test the zero hypothesis, dividing the pair differences is symmetric around zero \( ri = p_{ij} - \pi_{ij} \), \( H_0: F(-r) = 1 - F(r) \), a Wilcoxon pair test will be used.

g. Calculation of empirical logits \( h_{ij} = \ln \left( \frac{p_{ij}}{\pi_{ij}} \right), j = 1, 2, \ldots, J - 1 \).

h. Visualization of empirical and theoretical logits for individual web categories except for the reference one.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgments**

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0776/18 and VEGA-1/0821/21.

**References**

[1] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowl. Inf. Syst. 1 (1999) 5–32.

[2] B. Liu, Web data mining: exploring hyperlinks, contents, and usage data, 2011. doi:10.1007/978-3-642-19460-3.

[3] M. Munk, A. Pilkova, L. Benko, P. Blažeková, P. Švec, Pillar 3: pre-processed web server log file dataset of the banking institution, Data Br. (2021) Submitted for publication.

[4] M. Munk, A. Pilková, L. Benko, P. Blažeková, P. Švec, Web usage analysis of pillar3 disclosed information in turbulent times, Expert Syst. Appl. 185 (2021) 115503 (in press), doi:10.1016/j.eswa.2021.115503.

[5] M. Munk, A. Pilková, M. Drlik, J. Kapusta, P. Švec, Verification of the fulfillment of the purposes of Basel II, pillar 3 through application of the web log mining methods, Acta Univ. Agric. Silvic. Mendelianae Brum. 60 (2012).

[6] M. Munk, J. Kapusta, P. Švec, Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor, Procedia Comput. Sci. (2010) 2273–2280, doi:10.1016/j.procs.2010.04.255.

[7] M. Munk, J. Kapusta, P. Švec, M. Turčaní, Data advance preparation factors affecting results of sequence rule analysis in web log mining, E a M Ekon, a Manag 13 (2010).

[8] M. Munk, L. Benko, M. Gangur, M. Turčaní, Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining, E-M Ekon, a Manag 18 (2015) 144–159 doi:10.15240/1001/2015-3-013.

[9] M. Munk, M. Drlik, M. Vrábelová, Probability Modelling of Accesses to the Course Activities in the Web-Based Educational System, Comput. Sci. Its Appl. - Iccsa (2011) 485–499 Pt V, 2011.

[10] M. Munk, M. Drlik, Analysis of stakeholders’ behaviour depending on time in virtual learning environment, Appl. Math. Inf. Sci. 8 (2014) 773–785.

[11] M. Munk, M. Vrábelová, J. Kapusta, Probability modeling of accesses to the web parts of portal, Procedia Comput. Sci. 3 (2011) 677–683, doi:10.1016/j.procs.2010.12.113.