Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 1 (2010) 2273–2280

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Science, ICCS 2010

Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor

Michal Munk[a], Jozef Kapusta[a], Peter Švec[a*]

[a]*Constantine the Philosopher University in Nitra, Department of Informatics, Tr. A.Hlinku 1, 949 74 Nitra, Slovakia*

Abstract

Presumptions of each data analysis are data themselves, regardless of the analysis focus (visit rate analysis, optimization of portal, personalization of portal, etc.). Results of selected analysis highly depend on the quality of analyzed data. In case of portal usage analysis, these data can be obtained by monitoring web server log file. We are able to create data matrices and web map based on these data which will serve for searching for behaviour patterns of users. Data preparation from the log file represents the most time-consuming phase of whole analysis. We realized an experiment so that we can find out to which criteria are necessary to realize this time-consuming data preparation. We aimed at specifying the inevitable steps that are required for obtaining valid data from the log file. Specially, we focused on the reconstruction of activities of the web visitor. This advanced technique of data preprocessing belongs to time consuming one. In the article we tried to assess the impact of reconstruction of activities of a web visitor on the quantity and quality of the extracted rules which represent the web users' behaviour patterns.

Keywords: data preprocessing, data cleaning, identification of sessions, reconstruction of activities of a web visitor, web log mining, evaluation

1. Introduction

Good quality data are a prerequisite for a well-realized data analysis. If there is "junk" at the input, the same will be at the output, regardless of the method for knowledge extraction used. This applies even more in the area of web log mining, where the log file requires a thorough data preparation. As an example we can present the usage analysis, where we are aimed at finding out what our web visitors are interested in.

For this purpose we can use:

- survey sampling – we find out answers to particular items in the questionnaire and a visitor of our site knows that he/she is the object of our survey,
- web log mining – we analyzed the log file of the web server, which contains information on accesses to the pages of our web, and the visitor does not know that he is the object of our survey [1].

While in case of the survey sampling we can provide good quality data using a reliable and valid measuring procedure for their acquisition, in case of the web log mining we can provide them through good preparation of data

* Corresponding author. Tel.: +421-37-6408-675; fax:+421-37-6408-556.
E-mail address: psvec@ukf.sk.

1877-0509 © 2010 Published by Elsevier Ltd.
doi:10.1016/j.procs.2010.04.255

from the log file. In its standard structure called Common Log File it records each transaction, which was executed by the browser at each web access. Each line represents a record with the IP address, time and date of the visit, accessed object and referenced object. In case that we will use its extended version, we can record user browser version, the so-called User-Agent. In such data we follow series – sequences in visiting individual pages by the user, who is, under certain condition, identified by the IP address. In sequences we can look for users behaviour patterns. For this purpose it is the best way to use sequence rule analysis, the aim of which is to extract sequence rules. By means of these rules sequences of visits of various web sections by the user are predicted. This method was deduced from association rules and can serve as an example of the method making provision for Internet peculiarities.

The data preparation itself represents the most time consuming phase of the web page analysis [2]. The aim of the article is to find out using an experiment to what measure it is necessary to execute data preparation for web log mining and determine inevitable steps for gaining valid data from the log file. More accurately, our aim is to assess the impact of reconstruction of the activities of a web visitor on the quantity and quality of the extracted rules which represent the web users' behaviour patterns.

The rest of this paper is structured as follows. In section 2 we summarize particular problems which occur in data preprocessing for web log mining. Detailed methodology of our research is described in section 3, as the description of log files prepared in different levels of data preprocessing. End the end of this section we postulate what is the impact of reconstruction of activities of a web visitor for quantity and quality of discovered users' behavior patterns. Finally in section 4, we provide summary of experiment results and in section 5 we evaluate them.

2. Related work

Log file of the web server is a source of anonymous data about the user. These anonymous data represent also the problem of unique identification of the web visitor. Reconstruction of activities of each visitor is demanding. Currently it is common that several users share a common IP address, whether they are situated under a certain NAT (Network Address Translation), or proxy equipment. Authentication mechanisms can facilitate identification of the user, however, their usage is undesirable due to privacy protection [3]. Another problem which web log mining should face, are crawlers of various search engines, which browse through the whole web, mostly recursively and successively. Detection of crawlers is possible either based on their identification by means of the User-Agent field, or IP address by their comparison with the www.robotstxt.org database. This database need not contain data on all crawlers, however, those minority ones represent a neglectable number. Another method of identification is to find, whether crawler accessed to the file robots.txt or not [4]. Based on the access to this file we can unambiguously identify the crawler even if his User-Agent array is incorrectly set.

One of the possibilities how to differentiate individual visitors is to do it on the various versions of the Internet browser [1]. We can expect that if there exist several accesses from a single IP address with various versions of the browser or operating system, there is not only one user [5]. Cooley et. al. [1] also assume that if an access to the page, which is not accessible from the previous page, has been recorded, such an access can be accessed as the one by other user. This observation, however, is not unequivocal, since the user can run records on his/her favourite items and thus access also such subpages, which are not referenced from the previously accessed page [3].

Individual visitors can be differentiated also based on the identification of sessions. The aim of sessions identification is to divide individual accesses of each user into separate relations [1]. These relations can be defined in various ways. Session can be defined as a series of steps, which lead to the fulfilment of a certain task [6], or as a series of steps, which lead to the reaching of a certain goal [7]. The simplest method is to consider a session to be a series of clicks for a certain period of time, e.g. 30 minutes [3]. A real value for session can be derived based on empirical data.

Another problem of web log mining is a reconstruction of activities of a web visitor. Taucher and Greenberg [8] proved that more than 50 % of accesses on web are backward. This is the beginning of the problem with the browser´s cache. In backward, no query to web server is executed, so there does not exist any record of it in the log file. One of the solutions of this problem is path completion, through which we add these missing records to the log file [1].

In our experiment we tried to find out what steps were necessary when using sequence rule analysis.

3. Experiment

3.1. Data preprocessing

The data preparation for the needs of our experiment consisted of the following steps: data cleaning, identification of sessions and reconstruction of activities of a web visitor.

3.1.1. Data cleaning

The first step in the log file adjustment was **cleaning the file** from useless data. Under useless data we understand mainly the lines of the log file, in which are recorded requests for images, styles and scripts or other files, which can be inserted into the page. This part is the most simple from the whole data preparation process, since it consists of only filtration of the data, which do not comply with the selected template. A file of raw data of the log file resulted from this step. The pages are accessed also by **crawlers** of various **search engines**, which proceed in a different way than a common visitor does. By means of a simple detection of these crawlers and their deletion from the log file the resulting file with accesses only from standard visitors was obtained.

3.1.2. Identification of sessions

Another step in the advance preparation of data was **identification of sessions**. Identification of sessions of the user allows us to eliminate NAT and proxy devices, as well as identify various users alternating behind one computer. From our point of view sessions were identified as a delimited series of clicks realized in the defined period of time. In spite of the recommended 30-minute-long time window we chose the 10 minute time window with regard to the variable avg. time on site obtained by means of the Google Analytics tool, which represents average time of the user on our web page. While the current steps were the concern of simple programs, which sequentially scanned Common Log File, upon identification of the sessions it was necessary to enter all data from the file into the database, with which our application for the advance preparation of data for sequence rule analysis cooperated. After application of the algorithm ensuring identification of sessions to the file cleaned from crawlers we obtained file (File 1) - identification of sessions. One of the methods of identification of users (hiding behind various NAT devices or proxy servers) is their definition based on the **used web browser**, i.e. records from identical IP address were more specifically divided into individual sessions as to the used browser. This way we can specify also the sessions of users from Internet cafés, computer classrooms, etc., where several users alternate behind one computer, and we assume that not all of them use the same web browser. The result of this modification was the file with a closer division of users sessions, i.e. (File 2) - identification of sessions/agent.

3.1.3. Reconstruction of activities of a web visitor

Another problem upon searching for the users behaviour patterns seems to be the analysis of the backward path, or **reconstruction of activities of a web visitor**. Reconstruction of activities is focused on retrograde completion of records on the path went through by the user by means of a **back button**, since the use of such button is not automatically recorded into the Common Log File. A sitemap has a great importance for retrograde completion of the path. We can find in it information on the existence of a link among pages, i.e. if a hyperlink from one page to another exists. The sitemap was obtained for the needs of our analysis by means of Web Crawling application implemented in the used Data Miner. Having lined up the records according to the IP address we searched for some linkages between the consecutive pages. A sequence for the selected IP address can look like this: A→B→C→D→X. Based on the sitemap the algorithm in our example can find out that there does not exist the hyperlink from the page D to our page X. We thus assume that this page was accessed by the user by means of using a Back button from one of the previous pages. Through a backward browsing we then find out, on which of the previous pages exists a reference to page X. In our sample case we can find out that if there does not exist a hyperlink to page X from page C, if C page is entered into the sequence, i.e. the sequence will look like this: A→B→C→D→C→X. Similarly, we shall find that there does not exist any hyperlink from page B to page X and add it into the sequence, i.e. A→B→C→D→**C**→**B**→X. Finally algorithm finds out that page A contains hyperlink to page X and after the termination of the backward path analysis the sequence will look like this:

A→B→C→D→C→**B**→**A**→X. It means then, that the user used Back button in order to transfer from page D to C, from C to B and from B to A. After the application of this method to file (File 2) we obtained file (File 3) - identification of sessions/agent/path. For the needs of our experiment all these steps were taken using independent applications, i.e. we can simply say that a separate application was created for each step.

3.2. Research methodology

Experiment was realized in several 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. Creation of data matrices – from the log file (information of accesses) and sitemaps (information on the web contents).
3. Data preparation on various levels:
   3.1. with an identification of sessions (File 1 - identification of sessions),
   3.2. with an identification of sessions an agent allowance (File 2 - identification of sessions/agent),
   3.3. with an identification of sessions with making provisions for the agent and completing the paths (File 3 - identification of sessions/agent/path).
4. Data analysis – searching for behaviour patterns of web users in individual files.
5. Understanding the output data – creation of data matrices from the outcomes of the analysis, defining assumptions.
6. Comparison of results of data analysis elaborate on various levels of data preparation from the point of view of quantity and quality of the found rules – patterns of behaviours of users upon browsing the web.

We articulated the following assumptions:

1. we expect that completion of paths will have a significant impact on the quantity of extracted rules,
2. we expect that completion of paths will have a significant impact on the quality of extracted rules in terms of increasing the portion of inexplicable rules,
3. we expect that completion of paths will have a significant impact on the quality of extracted rules in the term of their basic measures of the quality.

4. Results

Users´ accesses to individual web sections of the portal were observed in the course of one week.

Table 1. Number of accesses and sequences in particular files

	Count of web accesses	**Count of costumer's sequences**	**Count of frequented sequences**	**Average size of costumer's sequences**
File 1	35374	8875	69	4
File 2	35374	9259	63	4
File 3	60087	9259	80	6

Having completed the paths (Tab. 1) the number of records increased by almost 70% and the average length of sequences increased from 4 to 6. On the other hand, upon making provision for the used browser (agent) when identifying sessions we can follow only about 4 % growth of visits (costumer's sequences).

The analysis (Tab. 2) resulted in sequence rules, which we obtained from frequented sequences fulfilling their minimum support (in our case min s = 0.03). Frequented sequences were obtained from identified sequences, i.e. visits of individual portal users during one week.

Table 2. Discovered sequence rules in particular files

Body	==>	Head	File 1	File 2	File 3
(a49)	==>	(a180)	1	1	1
(a49)	==>	(a358)	1	1	1
(a49)	==>	(a369)	1	1	1
(a49)	==>	(a369), (a49)	0	0	1
⋮	==>	⋮	⋮	⋮	⋮
(c3)	==>	(a134)	1	0	0
(c3)	==>	(a178)	1	1	0
(c3)	==>	(a180)	1	1	1
(c3)	==>	(a180), (a178)	1	1	0
(c3), (a180)	==>	(a178)	1	1	0
⋮	==>	⋮	⋮	⋮	⋮
(c7)	==>	(a49)	1	1	0
(c7)	==>	(a64)	1	1	1
(c7)	==>	(c3)	0	0	1
(c7)	==>	(c6)	1	0	1
Count of derived sequence rules			45	40	84
Percent of derived sequence rules (Percent 1's)			47.4	42.1	88.4
Percent 0's			52.6	57.9	11.6
Cochran Q Test			Q = 55.26984, df = 2, p < 0.000000		
Kendall Coeff. of Concordance			0.29089		

There is a high coincidence between the results (Tab. 2) of sequence rule analysis in terms of the portion of the found rules in case of files without completing paths (File 1, File 2). The most rules were extracted from file with completing paths, concretely 84 were extracted from the file (File 3), which represents over 88 % of the total number of found rules.

Based on the results of Q test (Tab. 2), the zero hypothesis, which reasons that the incidence of rules does not depend on individual levels of data preparation for web log mining, is rejected at the 1 % significance level.

Kendall´s coefficient of concordance represents the degree of concordance in the number of the found rules among examined files. The value of coefficient (Tab. 2) is 0.29, while 1 means a perfect concordance and 0 represents discordancy. Low value of coefficient confirms Q test results.

Table 3. Homogeneous groups of examined files

File	Mean	1	2
File 2	0.421	****	
File 1	0.474	****	
File 3	0.884		****

From the multiple comparison (Tukey Unequal N HSD test) (Tab. 3) a single homogenous group consisting of files File 1 and File 2 was identified in term of the average incidence of the found rules. Statistically significant differences on the level of significance 0.05 in the average incidence of found rules were proved among File 3 and the remaining ones.

Completion of paths has an important impact on the quantity of extracted rules (File 3 vs. File 1, File 2). On the

contrary, making provisions for the used browser upon identifying sessions has no significant impact on the quantity of extracted rules (File 1, File 2).

We require from association rules that they be not only clear but also useful. Association analysis produces three elementary kinds of rules [9]:

- utilizable (useful, beneficial),
- trivial,
- inexplicable.

In our case upon sequence rules it is useless to consider trivial rules. We will differentiate only the utilizable and inexplicable rules.

The portion of inexplicable rules (Tab. 4) is approximately over 25 % higher in file with completing paths (File 3) like in files without completing paths (File1, File 2).

Table 4. Crosstabulations: 2 by 2 Tables – Rule x File

Rule\File	**B1**	**B3**	**Rule\File**	**B2**	**B3**
Utilizable	43	59	Utilizable	38	59
	95.6 %	70.2 %		95.0 %	70.2 %
Inexplicable	2	25	Inexplicable	2	25
	4.4 %	29.8 %		5.0 %	29.8 %
Total	45	84	Total	40	84
	100 %	100 %		100 %	100 %
Phi-square	**0.08798**		**Phi-square**	**0.07866**	

Phi-square represents the degree of relationship between two dichotomy variables (Rule, File). The value of coefficient (Tab. 4) is approximately 0.08 in both cases, while 1 means perfect relationship and 0 no relationship. The portion of inexplicable rules is independent from reconstruction of activities of a web visitor.

Completion of paths has impact on increasing portion of inexplicable rules, but this increase of inexplicable rules is not significant.

Quality of sequence rules is assessed by means of two indicators:

- support,
- confidence.

Results of the sequence rule analysis showed differences not only in the quantity of the found rules, but also in the quality.

Kendall´s coefficient of concordance represents the degree of concordance in the support of the found rules among examined files. The value of coefficient (Tab. 5) is approximately 0.36, while 1 means a perfect concordance and 0 represents discordancy.

From the multiple comparison (Tukey Unequal N HSD test) (Tab. 5) a single homogenous group consisting of all files (File 1, File 2, File 3) was identified in term of the average support of the found rules. Statistically significant differences were not proved in support of the found rules among examined files.

Table 5. Homogeneous groups for support of derived rules

Support	**Mean**	**1**
File 2	3.032	****
File 1	3.161	****
File 3	3.207	****
Kendall Coeff. of Concordance		**0.3633**

There were demonstrated differences in the quality in terms of confidence characteristics values of the discovered rules among individual files. The coefficient of concordance values (Tab. 6) is almost 0.17, while 1 means a perfect concordance and 0 represents discordancy.

From the multiple comparison (Tukey Unequal N HSD test) (Tab. 6) a single homogenous group consisting of files File 1 and File 2 was identified in term of the average confidence of the found rules. Statistically significant differences on the level of significance 0.05 in the average confidence of found rules were proved among File 3 and the remaining ones.

Results (Tab. 6) show that the largest degree of concordance in the confidence is among the rules found in the file with identification of sessions without allowance (File 1) and with allowance of the agent (File 2). On the contrary, discordancy is among file with completing paths (File 3) and the remaining files (File 1, File 2).

Table 6. Homogeneous groups for confidence of derived rules

Confidence	**Mean**	**1**	**2**
File 2	37.545	****	
File 1	37.959	****	
File 3	41.831		****
Kendall Coeff. of Concordance		**0.1650**	

Completing the paths has a substantial impact on the quality of extracted rules (File 3 vs. File 1, File 2). On the contrary, allowing the used browser upon identifying sessions has not any significant impact on the quality of extracted rules (File 1, File 2).

5. Conclusions

Z. Yang et al. [10] in their research described factors influencing the quality of web portals and considered well organized messages and relevant contents to be the most important factors of usability. Web portals do not represent only a source of information for clients, but also a substantial source of data, from which we can obtain knowledge on visitors of our web portal with the aim to optimize, analyze the visit rate, personalize the portal [11-14], etc..

Data on the use of the portal can be obtained by monitoring of the log file of the web server. We are able to create data matrices from accesses and the web map, which will serve for searching for behaviour patterns of users. An advantage of an analysis of the log file of the web server comparing to, for example, a selective detection consists in the fact that the visitor does not know that he is the object of investigation. On the other hand, a disadvantage is the mere preparation of data from the log file, which represents the most time-consuming phase of the analysis of the web page [2]. The experiment was realized with the aim to find out to which measure it is necessary to realize this time-consuming preparation of data and we aimed at specifying the steps inevitable for obtaining valid data from the log file. Specially, we focused on the reconstruction of activities of the web visitor. This advanced technique of data preprocessing belongs to time consuming and less using. In spite of that over 50 % of accesses on web are backward [8]. We tried to assess the impact of reconstruction of activities of a web visitor on the quantity and quality of the extracted rules which represent the web users' behaviour patterns.

The first assumption concerning the reconstruction of activities of a web visitor and its impact on quantity was fully proved. Completing paths was found crucial in the data preparation for web log mining. Specifically, it was proved that completing the paths has a significant impact on the quantity of extracted rules (File 3 vs. File 1, File 2). Statistically significant differences in the average incidence of found rules were proved among File 3 and the remaining ones.

On the contrary, the second assumption concerning the reconstruction of activities of a web visitor and its impact on quality in terms of increasing the portion of inexplicable rules was proved only partially. Completion of paths had impact on increasing portion of inexplicable rules, but this increase of inexplicable rules was not significant.

The third assumption concerning the reconstruction of activities of a web visitor and its impact on quality of extracted rules in term of their basic measures of quality was also fully proved. It was similarly proved that path

completion has a significant impact on the quality of extracted rules (File 3 vs. File 1, File 2). Statistically significant differences in the average confidence of found rules were proved among File 3 and the remaining ones.

Besides, it was showed that allowing the used browser upon identifying sessions has neither significant impact on quantity nor quality of extracted rules (File 1, File 2).

Searching for rules by means of sequence rule analysis is closely connected with certain steps of data preparation. It was proved that completing the paths is very important, however, it depends to a great degree on the correct identification of individual sessions of the portal visitors. There exist a large number of models for the identification of users sessions [15-19]. There exists also a method, which expressly identifies these sessions. This method is called additional programming of an application, which creates web logs. We thus obtain a more sophisticated solution, by means of which on the one hand we are able to identify, expressly and confessedly, each visitor´s session, but on the other hand we lose the general method of web log processing.

Path completion also depends on the topicality of the sitemap, which can be modified too quickly so that we could use the offline method of web log analysis.

References

1. R. Cooley, B. Mobasher and J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information System, 1999, Springer-Verlag, Vol. 1, ISSN 0219-1377.

2. P. Berka, Dobývání znalostí z databází. Praha, 2003. ISBN 80-200-1062-9.

3. B. Berendt and M. Spiliopoulou, 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.

4. A. G. Lourenco, O. Belo, Catching web crawlers in the act. Proceedings of the 6th international Conference on Web Engineering, 2006, ICWE '06, Vol. 263, ACM, New York, NY, pp. 265-272. ISBN 1-59593-352-2.

5. P. Pirolli, J. Pitkow and R. Rao, Silk from a sow's ear: Extracting usable structures from the Web. Proc. of 1996 Conference on Human Factors in Computing Systems (CHI-96), 1996, Vancouver.

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

7. M. Chen, J.S. Park, and P.S. Yu, Data mining for path traversal patterns in a web environment. ICDCS, 1996, pp. 385–392. ISBN 0-8186-7398-2.

8. L. Taucher and S. Greenberg, Revisitation patterns in world wide web navigation. Proc. of Int. Conf. CHI'97, 1997, Atlanta.

9. I. Stankovičová, Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner. Informační a datová bezpečnost ve vazbě na strategické rozhodování ve znalostní společnosti, 2009, pp. 1-9. ISBN 978-807318-828-3.

10. Z. Yang, S. Cai, Z. Zhou and N. Zhou, Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information & Management, 2005, Volume 42, Issue 4, pp. 575-589. ISSN 0378-7206.

11. M. Drlík, User Interface Adaptation. ASIS 2008, 2008, pp. 41-52. ISBN 978-80-8094-400-1.

12. J. Skalka, User Modelling by Neural Networks. ASIS 2008, 2008, pp. 135-141. ISBN 978-80-8094-400-1.

13. C. Klimeš and Z. Balogh, Fuzzy Adaptation Model. ASIS 2009, 2009. pp. 51-58. ISBN 978-80-8094-593-0.

14. M. Turčáni, E-learning v prostredí LMS Moodle s podporou adaptívnych hypermédií. ERIE - Efficiency and Responsibility in Education, 2009, pp. 79. ISBN 978-80-213-1938-7.

15. B. Hay, G. Wets and K. Vanhoof, Web usage mining by means of multidimensional sequence alignment methods. WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profile, Lecture Notes in Computer Science, 2003, Vol. 2703, Springer, Berlin/Heidelberg, pp. 50–65.

16. B. Hay, G. Wets and K. Vanhoof, Segmentation of visiting patterns on Web sites using a sequence alignment method. Journal of Retailing and Consumer Services, 2003, Vol. 10, No. 3, pp. 145–153. ISSN 0969-6989.

17. F. Masseglia, D. Tanasa and E.B. Trousse, Web usage mining: Sequential pattern extraction with a very low support. Advanced Web Technologies and Applications, Lecture Notes in Computer Science, 2004, Vol. 3007, pp. 513–522. ISBN 978-3-540-21371-0.

18. S. Oyanagi, K. Kubota and A. Nakase, Mining WWW access sequence by matrix clustering. Mining Web Data for Discovering Usage Patterns and Profiles, Lecture Notes in Computer Science, 2003, Vol. 2703, Springer, Berlin/Heidelberg, pp. 119–136. ISBN 978-3540203049.

19. S. Park, N.C. Suresh and B. Jeong, Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. Data & Knowledge Engineering, 2008, Vol. 65, pp. 512–543. ISSN 0169-023X.