Applied Mathematics & Information Sciences
An International Journal

Methodology Design for Data Preparation in the Process of Discovering Patterns of Web Users Behaviour

Michal Munk, Martin Drlík, Jozef Kapusta, Daša Munková

Constantine the Philosopher University in Nitra, Nitra 949 74 Slovak Republic

Received: 19 Oct. 2012, Revised: 10 Nov. 2012, Accepted: 9 Jan. 2013
Published online: 1 Feb. 2013

Abstract: Discovering of behaviour patterns of website visitors is one of the most common applications in web log mining. Based on the discovered users' behaviour patterns, it is possible to restructure or in combination with other knowledge personalize the examined website, portal or other web-based system. Data preparation represents the first inevitable step in the process of discovering users' behavioural patterns. In this paper we summarize the results of our previous research, where we carefully examined the relevance of individual steps of data preparation from a web server log file and virtual learning environment for further analysis. The aim of our experiments was to find out to what extent it is necessary to realize the time-consuming data preparation in the process of discovering patterns of behaviour of web users and to determine the inevitable steps to obtain reliable data from different types of log files. Considering the obtained results we propose a methodology for data preparation in the process of discovering patterns of web user behaviour based on the results of experiments we carried out. The research results showed, that in the case of systems providing sophisticated navigation options and a rigid structure of the content (which is characteristic for the most virtual learning environments), the paths completing is not an inevitable step in data preparation in the process of discovering patterns of web users' behaviour.

Keywords: Web log mining, data preparation, user behaviour, discovering patterns

1 Introduction

The most time-consuming stage in the process of knowledge discovery itself is the data preparation. This is especially true in the area of the knowledge discovery based on the use of Web (Web Log Mining, WLM). The problem in data preparation is not data transformation into the form required by analytical tools, but rather the quality - reliability of data.

In this paper we summarize the results of our previous research that we partially published in publications [1–5]. There, we evaluated the relevance of individual steps of data preparation from a web server log file and a virtual learning environment (Virtual Learning Environment, VLE) for sequence analysis.

We verified the importance of the possibilities of data preparation from the log file by means of experiments in which we were trying to find out which steps of the data preparation are important for a correct analysis of the portal with anonymous access [1–3] and virtual learning environment [4,5].

The results of experiments are very important for the portal/system, which is regularly analysed and modified, since they can prove correctness of the individual steps in data preparation, or through the identification of "useless" steps simplify the data preparation. In our experiments the discovered knowledge is represented by sequential rules. Discovering patterns of behaviour of web site visitors is one of the most common applications in web log mining.

Based on the discovered users' behaviour patterns, which are represented by sequential rules, it is possible to restructure or in combination with other knowledge personalize the examined portal, or system. The first studied data source is a web server log file [1–3], which in the case of portals presents the most widely used source of WLM area. The second data source is a log file of VLE system [4,5], namely Moodle system, that in comparison with the web server log file stores data of the use in its own structure organized in a relational database. In terms of knowledge discovery the sources can be considered as valid data sources of using the portal, or system, provided that the inevitable steps will be realized in the data preparation.

The aim of our experiments is to find out to what extent it is necessary to realize the time-consuming data

* Corresponding author e-mail: mmunk@ukf.sk

preparation in the process of discovering patterns of behaviour of web users and to determine the inevitable steps to obtain reliable data from the log file.

2 Related Work

VLM has received extensive attention because of its significant theoretical background and great application potential. Many web usage mining approaches and methods were surveyed in [6–9]. The last comprehensive surveys on WUM have been done by Koutri, Avouiris and Daskalaki [10] and Kosala and Blockeel [11].

Facca and Lanzi summarize recent developments in WUM research in their expert studies [12] and [13] 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 [14].

In the first phase, web log data are pre-processed in order to identify users, sessions, page views, and clickstreams [15].

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 [14] it consists of

1. data cleaning (for removing irrelevant references and fields, removing erroneous references, adding missing references due to caching mechanisms, etc.) [16, 17],
2. data integration (for synchronizing data from multiple server logs, integrating registration data, etc.) [18–20],
3. data transformation (for user-session identification [21, 22], path completion [23, 24], etc.),
4. and data reduction (for reducing dimensionality) [25, 26].

Pattern discovery of web usage mining which is the outcome of the proposed methodology is discussed in detail in [18], [27] and [28]. According to Song et al. [27] the behaviour pattern discovering is one of the WLM sub-areas that focus on finding out typical flow models of user actions from recorded events. This process is iterative and incremental and can be divided into four phases that are similar to the phases of VLM [29]. Another approach to behaviour pattern discovering based on ontology and sequence information was presented in [30].

Recently, a number of researches have been undertaken to analyse pros and cons of data mining for e-learning decision making [31, 32]. Exploratory analysis and WLM were used to explore students' behaviour in [33]. We can find other methods and applications of VLM methods used in distance learning in [34, 35] and [36]. Data pre-processing phase of web usage mining is described in [37–39] in detail. Koutri et al [10] prepared an excellent survey on web usage mining techniques for web-based adaptive hypermedia systems. We reviewed other applications of WLM in e-commerce [40, 41], in recommendation systems development [42, 43] or search engine development [44].

The above-mentioned papers and other explored references use several data mining methods, but do not deal with the similar methodology as is being described in our paper.

The available resources show that the structured methodology for data preparation in the process of discovering patterns of web users' behaviour is not systematically elaborated. We found the only reference to similar methodology as proposed in our paper in [45]. Unlike our proposal this methodology is designed for the bank sector and marketing.

3 Research Methodology

We used a uniform research methodology in all experiments. When examining the impact of data preparation on the quantity and quality of the obtained knowledge we took the following steps:

1. Data acquisition from a web server log file, or from a virtual learning environment.
2. Data preparation on various levels - Creation of pre-processing data files on various levels of data preparation. In general, data preparation consists of the following tasks:
   (a) Data cleaning.
   Data cleaning involves removing useless data - requests for images, scripts and styles is an inevitable step. This is a starting point - a fundamental step in data preparation from the web server log file [1–3]. The result is a file of raw data, consisting of portal accesses. Data cleaning from the web server log file includes data cleaning from the accesses of search engines crawlers, where appropriate, data cleaning from the accesses of NAT/proxy devices.
   The object of our investigation is to find out to what extent these steps misrepresent the analysis results [1].
   Cleaning data from the log file of VLE system includes only removing the accesses of the user groups whose behaviour is not the object of our investigation. Like in the previous case, this step is the starting point in the data preparation from the log file of the virtual learning environment [4, 5].
   (b) Identification of users/sessions
   In the first chapter of the paper we introduced several techniques of identifying users/sessions. During data preparation from the web server log file, we focus on the most commonly used ones,

namely techniques for identification of users based on agent and identification of sessions based on time [2,3].

In the case of virtual learning environments, the aim is not to identify users, but the transaction/sequence of system users, i.e. in this case, we focus on identification of session based on time, while we experiment with various length of session timeout threshold, (STT) [5].

(c)Reconstruction of activities of web users.

For the reconstruction of the activities of web users, we use a current site map, thus we achieve a more accurate paths completion [2–5].

The aim of our research is to determine to which extent the paths completion affects the quantity and quality of the knowledge obtained from using the portals with anonymous access [2], and how the basic steps of data preparation affect the paths completion itself [3].

Similarly, the aim of our research is also to state the impact of paths completing of obtained knowledge upon the use of virtual learning environments [4].

3.Data analysis - searching for behaviour patterns of web users in individual files. During the search for behaviour patterns in the examined files, it is necessary to ensure that the rules of individual files were extracted under the same conditions.

4.Understanding output data - Design of the data file from the outputs of the analysis of individual files and a calculation of the basic characteristics of the examined files:

(a)the number of accesses,
(b)the number of customer/identified sequences,
(c)the number of frequented sequences,
(d)the average size/length of the identified sequences.

Based on these primary results it is possible to specify the assumptions.

5.Comparison of knowledge obtained from the examined files pre-processed on various levels of data preparation.

In the assessment of the obtained knowledge we focus not only on the quantity of extracted rules, but also on their quality. The quality of sequence rules is assessed by two indicators [46,47]:

(a)support,
(b)confidence.

Also in the evaluation of obtained knowledge we consider their applicability in practice. We require from sequence rules, as well as from association rules, to be not only understandable but also useful. In general, sequence analysis produces three types of rules [46]:

(a)useful,
(b)trivial,
(c)inexplicable.

Useful rules contain high quality information, trivial rules contain a result, which is widely known for the given area, and inexplicable rules cannot be explained and do not lead to any useful action [47]. At this stage cooperation with an expert through data from the application area is very important. In our case, in the evaluation of obtained knowledge from the web server log file, we collaborate with the portal creators and administrators, and in case of a log of VLE system with the system administrators and course creators.

Acquired knowledge is evaluated in terms of the quantity and quality of the discovered sequence rules - behaviour patterns of users through browsing the web in terms of:

(a)comparison of the proportion of the rules found in the examined files,
(b)comparison of the proportion of useful, or trivial and inexplicable rules in the examined files,
(c)comparison of values of the degree of support and confidence of the rules found in the examined files.

4 Research Result for Portal with Anonymous Accesses

Good quality data - reliable data are a prerequisite for a well-realized data analysis. The web server log file data contain unnecessary, irrelevant, inaccurate and incomplete information about the use of web. Unnecessary data are lines of log file, in which the requests for images, styles and scripts or other files that can be inserted into the page are recorded. Irrelevant data are lines of log file, in which the accesses of crawlers of various searching engines recorded not the accesses of users - web visitors whose behaviour is the object of the analysis. The inaccuracy of the data relates to the anonymous character of data. We consider the log file a source of anonymous data about the user in a way that we do not record his/her personal data or unambiguous identification. Incompleteness of data causes the browser's cache. In the web server log file, the paths records, which the user passed through the button back, are missing. Reliable data are ensured with quality data preparation from a log file in web log mining. We investigated through series of experiments which steps in the data preparation are inevitable to obtain reliable data from web server log file [1–3].

In the first experiment [1] we attempted to find out, how important are the basic steps of data preparation for the sequence analysis use, i.e. the aim was to evaluate the relevance of the basic steps of data preparation with the emphasis on data cleaning.

For the purposes of our experiment, we carried out data preparation in four different levels:

1.data cleaning from unnecessary data/requests for images, scripts and styles - raw data,
2.data cleaning from unnecessary data and crawlers' accesses,
3.data cleaning from unnecessary data, crawlers' accesses and NAT/proxy devices,

4.data cleaning from unnecessary data, crawlers‘ accesses and with identification of sessions based on time.

We expected that the different levels of data preparation would have a significant impact on the quantity of extracted rules as well as on their quality in terms of the basic quality characteristics.

The experiment was focused on the basic steps of data preparation and served rather in order to confirm the generally valid assumptions and to verify the research methodology. Despite that, the results of the experiment revealed several important facts.

It is interesting, that most of the rules were found just after the removal of various crawlers' records of searching engines. It was shown that the recursive character of portal browsing by crawlers could distort the results. However, in the data preparation in WLM process the exclusion of the crawlers search engines had no significant impact on the results of sequence analysis. Similarly, the elimination of NAT/proxy devices had no significant effect on the results.

On the contrary, the identification of users‘ sessions based on time had the expected impact on the accuracy of the analysis results. Just the identification of sessions seems to be the most important factor in the whole data preparation. By the identification of the sessions we just did not allow the inclusion of NAT/proxy device into the analysis, but we also eliminated the problem of "one computer more sessions," whereby a number of unrelated sequences were significantly reduced. For Internet cafes, libraries, classrooms and so on, it is specific that more than one anonymous user uses one computer, this fact caused the elimination of identifying sessions based on time. The discovered rules showed the same, only in the file results with the identification of the sessions, the inexplicable rules did not occur.

Data preparation can be further refined. For example, sessions can be specified and people with various agents (browsers) can access from a single IP address. This factor can specify the analysis results. One of the methods of identifying the 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 could be more specifically divided into individual sessions according to the used browser. This way we can also specify the sessions of users from Internet cafes, computer classrooms, etc., where several users alternate at one computer, and we assume that not all of them use the same web browser.

The important step in data preparation during the search for the users' behaviour patterns seems to be also the analysis of the backward path, or reconstruction of activities of a web visitor. The reconstruction of activities is focused on retrograde completion of records on the path created by the user by means of a back button, since the use of such button is not automatically recorded into the log file. In another experiment [2] we focused not only on accuracy of the sessions but also on the reconstruction of the activities of the web users.

For the needs of our experiment we cleaned data from the useless data and crawlers‘ accesses and then we carried out next data preparation on three different levels:

1.identification of sessions based on time,
2.identification of sessions based on time with an agent allowance,
3.identification of sessions based on time with making provisions for the agent and completing the paths.

We expected that completion of paths will have a significant impact on the quantity of extracted rules as well as on the quality of extracted rules in terms of their basic characteristics of the quality. Similarly, we expected that completion of paths will have a significant impact on increasing the portion of inexplicable rules.

The experiment results proved that completing the paths has a significant impact on obtained knowledge. Specifically, there are statistically significant differences in average incidence and reliability of the discovered rules between the file with completing the paths and without it. On the contrary, the assumption concerning the increased portion of inexplicable rules was proved only partially. It was proved that the paths completion has an impact on the increasing portion of the inexplicable rules but this increase is not statistically significant.

We expect that the correct paths completion depends on the basic steps of data preparation as data cleaning or identification of sessions. We verified to what extent these steps effect the paths completion in the last experiment [3] of series of experiments focusing on data preparation from the web server log file. The aim of the experiment was to find out to what extent it is necessary to execute the time consuming data preparation in the process of discovering behaviour patterns of the web users and to specify the inevitable steps for obtaining reliable data from the log file. The research is connected with the results from the previous experiments [1,2], where we evaluated the relevance of individual steps of data preparation by discovering patterns of the web users behaviour.

Data preparation for the needs of our experiment itself consisted of the following steps:

1.The first step in the log file adjustment was removing useless data from the file. The result of this step was a raw data file with accesses to the portal pages (raw data).
2.The pages are accessed also by crawlers of searching engines. By detection of these crawlers and their deletion from the log file we obtained the resulting file with accesses only from standard visitors (without crawlers).
3.Another step in the data preparation was identification of sessions based on time. In spite of the recommended 30-minute-long time window [48] we chose the 10 minute time window (STT) with regard

to the variable average time on site obtained by means of the Google Analytics tool, which represents average time of the user on our web page. After application of the algorithm ensuring identification of sessions to the file of raw data and the file cleaned from crawlers we obtained files A1 (raw data/identification of sessions) and B1 (without crawlers/identification of sessions).

4. One of the methods of identification of users 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 or operation system. The result of this modification were the files A2 (raw data/identification of sessions/agent) and B2 (without crawlers/identification of sessions/agent) with a closer division of users sessions.
5. The last step of data preparation was a reconstruction of activities of a web visitor focusing on retrograde completion of records on the path created by the user by means of a back button, since the use of such button is not automatically recorded into the log file. After application of this method on the files A2 and B2 we obtained files A3 (raw data/agent/identification of sessions/path) and B3 (without crawlers/identification of sessions/path).

The following scheme (Fig. 4.1) provides an illustrative overview of applying techniques for log file data and the subsequent creation of files XY, where X = A, B and Y = 1, 2, 3 for the needs of the experiment.

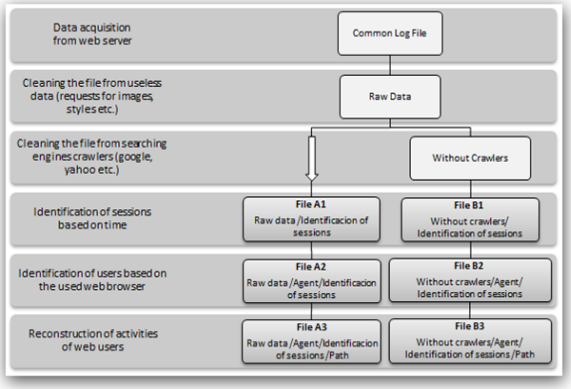


Fig. 1 Application of techniques of data preparation to web server log file.

Based on the results from the previous experiments [1, 2] we expected that cleaning the data from accesses of the crawlers searching services would not have a significant impact on the quantity and quality of the extracted rules in the term of their basic characteristics of the quality.

On the contrary, we expected that cleaning the data from the crawlers' accesses would have a significant impact on the reducing portion of the inexplicable rules. Further, we expected that the paths completion would have a significant impact on the quantity and also the quality of the discovered knowledge and allowing the used browser during identifying sessions would not have a significant impact on the analysis results. The question to what extent do the basic steps of data preparation affect the completion of paths remained unanswered.

The first assumption concerning the data cleaning was proved only partially. Specifically, data cleaning from the crawlers' accesses has a significant impact on the quantity of extracted rules only in case of files with completing paths (A3 vs. B3). It could be caused by the fact that the crawlers of different searching engines browse the web sequentially. If we apply the paths reconstruction into this search, the program will generate an amount of non-standard data. The impact on reduction of the portion of the inexplicable rules as well as the impact on the quality of the extracted rules in the term of their basic characteristics of the quality was not proved.

The second assumption was fully proved - the paths completion has shown itself to be the key in case of data preparation for WLM. Specifically, it was shown, that the paths completion has a significant impact on the quantity of extracted rules (X3 vs. X1, X2). Similarly, it was shown that the paths completion has a significant impact on the quality of extracted rules (X3 vs. X1, X2), regardless of the fact, whether the files are cleaned data from the crawlers accesses (BY) or not (AY).

The third assumption was also fully proved. Specifically, it was shown, that considering the used browser during identifying sessions has no impact on the quantity as well on the quality of the extracted rules (X1, X2), the same regardless of the fact, whether the files are cleaned data from the crawlers accesses (BY) or not (AY).

The data themselves are the requirement of each data analysis about the web use regardless its focus (analysis of the visit rate, restructuring, portal personalization etc.). The results of the analysis depend on the quality of the analysed data.

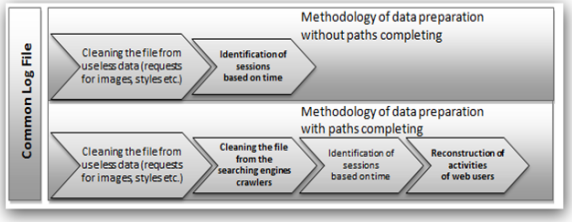


Fig. 2 Methodology design of data preparation based on the experiments' results for the log file in the CLF structure.

For this reason, in our experiment we aimed at specifying steps inevitable for data preparation from a web server log file. Based on the experiments conclusions, the scheme 4 depicts inevitable steps to be taken in order to obtain reliable data [1–3].

Based on the experiments results, we propose two alternatives for data processing of a log file 4, the first alternative is simpler and less time-consuming for the data preparation at the cost of a lower number and less accurate extracted knowledge. The second, more time consuming alternative offers reliable data, and thus also a large amount and better quality of knowledge, represented by the discovered rules. In both proposed methodologies of data preparation, the step of definition of sessions based on the used web browser, which had no impact on the quantity or quality of the extracted rules [2,3] is absent. In the simpler methodology, which presents an alternative without the reconstruction of the activities of a web visitor, the step of cleaning the file from the crawlers of searching services, which had an impact on the quantity of extracted rules only in case of files with the completion of the paths [3] is also absent.

5 Research Results for VLE

The last two experiments are aimed at specifying inevitable steps of data preparation for obtaining reliable data from the log file of the virtual learning environment, but are based on previous experiment results [1–3].

The aim of the first two experiments [4] was to evaluate the impact of identification of sessions and path completion on knowledge discovery represented by the behaviour patterns of students - users of a virtual learning environment. We cleaned data from the users‘ accesses whose role was different than of a student in the system and subsequently we carried out another data preparation on three different levels:

1. identification of sessions based on user ID and IP address,
2. identification of sessions based on user ID, IP address and time,
3. identification of sessions based on user ID, IP address, time and completing the paths.

Provided that we would identify the sessions based only on a user ID, the individual student course visits during the whole study time would join together into one session. In our case, the session presents the period of at least 13 weeks. The consequence of such defined sessions is the identification of a disproportionate number of frequented sequences and also the inexplicable rules and high values of the characteristics of the discovered rules.

In our case it is a modification of the previous experiments [1,2] for the virtual learning environment. Therefore, based on the previous findings we expected that the identification of sessions based on time would have a significant impact on the quantity of obtained knowledge, i.e. on reducing the portion of trivial and inexplicable rules and vice versa, completing the path on increasing the portion of useful rules. We also assumed that the investigated techniques of data preparation would have a significant impact on the quality of obtained knowledge, i.e. on reducing the value of the basic characteristics of discovered rules.

Surprising finding, regarding also the previous experimental results, was that the paths completion has significant impact neither on the quantity nor on the quality of obtained knowledge. Completing the paths had an impact on increasing portion of useful rules, but this increase of useful rules was not statistically significant.

On the other hand, similarly as in the case of data preparation from a web server log file, identification of sessions based on the time is crucial in case of log file data in VLE system. Identification of sessions based on the time has a significant impact on the quantity as well as on the quality of the obtained knowledge. The portion of trivial and inexplicable rules is dependent on the identification of sessions based on time. The identification of sessions has a significant impact on the reduction of the portion of trivial and inexplicable rules, while the portion of useful rules remains preserved.

The aim of the second experiment [5] from the experiments focusing on data preparation from a log file of VLE system was to evaluate the impact of the length of session timeout threshold (STT) during identifying sessions based on time on the quantity and quality of the obtained knowledge.

We cleaned data from the users‘ accesses whose role was different than of a student in the system and subsequently we carried out another data preparation with a different STT length:

1. identification of sessions based on 15-minute STT,
2. identification of sessions based on 30-minute STT,
3. identification of sessions based on 60-minute STT,
4. identification of sessions based on 15-minute STT and the paths completion,
5. identification of sessions based on 30-minute STT and the paths completion,
6. identification of sessions based on 60-minute STT and the paths completion.

The following scheme 5 provides an illustrative overview of the application of individual techniques to log file of a virtual learning environment and the subsequent creation of files XY, where X = A, B and Y = 1, 2, 3 for the needs of the experiment.

We expected that the identification of sessions based on shorter STT would have a significant impact on the quantity of the obtained knowledge i.e. on reducing portion of trivial and inexplicable rules and on the quality in terms of the basic quality characteristics of discovered rules.

Despite the previous findings [4], we assumed that the path completing in combination with various lengths of

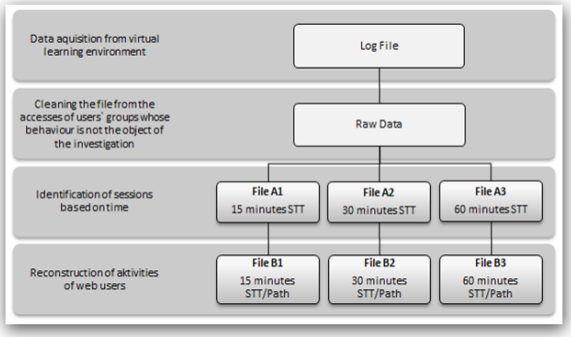


Fig. 3 Application of techniques of data preparation to log file of VLE system.

time window upon identifying sessions would have a significant impact on the quantity of the obtained knowledge in terms of increasing portion of useful rules and the basic quality characteristics of discovered rules.

The assumption concerning the identification of sessions based on time and its impact on quantity of extracted rules was fully proved. Specifically, it was proved that the length of STT has a significant impact on the quantity of extracted rules. Statistically significant differences in average incidence, support and confidence of discovered rules were proved among files X1, X2, X3 regardless of the fact, whether the files are with the path completion (BY) or not (AY). The portion of trivial and inexplicable rules is dependent on the STT length. Identification of sessions based on shorter STT has an impact on reducing portion of trivial and inexplicable rules as well as on the quality of discovered rules in term of the basic characteristics of quality.

On the other hand, it was showed that the completion of paths has neither significant impact on the quantity nor on the quality of the extracted rules (AY, BY). Paths completion has no significant impact on increasing portion of useful rules. While the completion of the path in combination with the identification of sessions based on 60-minute STT has a significant impact on the quantity of obtained knowledge, however, only the portion of trivial and inexplicable rules increased. Path completion with inappropriate identification of STT may cause an increase of trivial and inexplicable rules. Results show the highest degree of concordance in the support and confidence among the rules found in the file without completion of paths (AY) and in corresponding file with the completion of paths (BY). The assumption of an impact of paths completion on obtained knowledge was not proved.

Identification of sessions based on the time is crucial in data preparation from a log file of a virtual learning environment [4], as suggested by the previous findings in relation to data preparation from web server log file [2,3]. However, the correct estimation of STT length upon identifying sessions based on time is important. The choice of a too large time window (STT) could lead to the increasing of trivial and inexplicable rules while in combination with paths completion, this increase could even be much more significant [5]. Surprising finding, regarding also the previous experimental results [2,3], was that the reconstruction of activities of VLE system users has no significant impact on quantity and quality of obtained knowledge [4]. For comparison, after completing the paths in case of web server log file, the number of records increased by almost 70% [2].

On the contrary, in case of log file of VLF system only by about 7% [4]. We assume that this difference is caused by the rigid structure of the e-learning course and more sophisticated hierarchical menu. However, we see the main causes in the existence of breadcrumbs, which is available in any e-learning course of the virtual learning system Moodle that eliminates the use of back button in browsing the courses. The research results indicate that the data preparation through educational context can be reduced to reconstruction of the activities of system users. We assume that in case of a system providing sophisticated navigation options and a rigid structure of the content (which is characteristic not only for most virtual learning environments but also for other systems), paths completing is not an inevitable step in data preparation in the process of discovering patterns of web users behaviour.

Conclusions

The requirement of data analysis is data itself. The quantity and quality of obtained knowledge depends on reliability of the analysed data. On the contrary, data preparation presents the most time-consuming stage of the process of knowledge discovery. Our aim was to find out to what extent it is necessary to carry out the most time-consuming data preparation in the process of discovering patterns of web users' behaviour and to specify the steps inevitable for obtaining reliable data from the log file. For this purpose, series of experiments focusing on data preparation from the web server log file in standard CLF structure were carried out. We investigated which steps of data preparation are important for a proper analysis of the portal with anonymous access and virtual learning environment. The contribution of our research is a proposal of a methodology and recommendations for obtaining reliable data from the log file in the process of discovering patterns of web users' behaviour. Based on the results of our experiments, we proposed two alternatives for data processing of the web server log file. The first one represents a simpler and less time-consuming data preparation at the cost of a lower number and less accurate extracted knowledge.

The second one, more time-consuming alternative with paths completing offers more reliable data, and thus

also a larger amount of better quality knowledge. The research results showed, that in the case of systems providing sophisticated navigation options and a rigid structure of the content (which is characteristic for most virtual learning environments), paths completing is not an inevitable step in data preparation in the process of discovering patterns of web users‘ behaviour.

The proposed methodologies and recommendations for obtaining reliable data from the log file were applied into the process of the web users‘ behaviour modelling in dependence on time [49,50]. We assume that in both experiments particular web pages as well as visitors can be divided into logical categories. We deal with the probabilities of accesses of particular groups of visitors to the individual web pages depending on the day of the week and on the hour of the particular day. These probabilities were estimated for each group of visitors using multinominal logit model [51,52].

This model was applied on data about the portal use with anonymous access and virtual learning environment. The obtained knowledge was used as a base in planning portal/system maintenance.

Identification of sessions based on time is crucial in case of data preparation from the web server log file and in case of log file of VLE system. In the case of portal with anonymous access it was showed that the path completing is very important; however, it depends on the correct identification of individual sessions of the portal visitors. We know that there is a large number of models and methods for identification of users’ sessions. Therefore we plan to prepare further experiments focused on using more precise user session identification methods.

One of these methods is for example an additional programming of an application which creates web logs. Thus, we obtain a more sophisticated solution, by means of which we are able to unambiguously and indisputably identify each visitor’s session on one hand, but on the other hand we lose the general method of web log processing. We additionally programmed this functionality into our analysed portal. Our further goal is to analyse various parameters of individual methods of identification of sessions compared with the reference direct identification.

Another interesting method that we consider to use in our future experiments deals with the comparison of presented user session identification method with other methods based on cookies [53].

We put special emphasis on the navigation-driven heuristic methods. These methods are based on the idea that each website contains content pages, navigation pages and auxiliary pages. Content pages are web pages where the visitor can find desired information. These pages represent targets of visitors’ searching and browsing activity. Other web pages (navigation and auxiliary) are important for successful navigation to the content pages. The user session can be defined as the path over several navigation and auxiliary web pages to the desired content page. User identification is based on two methods - the Reference length method and Forward-Reference method [14,54,55].

Another problem of user session identification based on the proposed methodology should be mentioned. The path completion depends on the topicality of the sitemap which can change too quickly for us to be able to use the offline method of web log analysis. Our further research will be devoted to dynamic analysis of web logs. We will concentrate on methods for extracting sitemap from the log file according to [23–25,56].

Our aim is to reduce time necessary for the pre-processing of these logs and at the same time to increase the accuracy of these data.

At present, we solve a project focusing on verification of the fulfilment of the purposes of Basel 2, Pillar 3 - market discipline during the recent financial crisis [57]. The aim of the project is to analyse the stakeholder‘s interest in mandatory disclosure of financial information by a commercial bank by means of advanced methods of web log mining and to find out whether the purposes of Basel 2, Pillar 3 have been fulfilled. The output of the project will be a verification of the assumptions related to the purposes of Basel 3 by means of the web mining methods and a formulation of recommendations for possible reduction of mandatory disclosure of information under Basel 2 and 3. On the other hand, the output of the project will be a verification of our methodology of data preparation in this specific domain of banking sector. For this purpose, a cooperating commercial bank provided the data about the use of their portal from the year 2008. We will have an opportunity to verify our existing findings on large data sets.

References

[1] M. Munk, Počítačová analýza dát, UKF, Nitra (2011).
[2] M. Munk, J. Kapusta and P. Švec, Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor. Procedia Computer Science. **1**, Elsevier, 2273-2280 (2010).
[3] M. Munk, J. Kapusta, P. Švec and M. Turčáni, Data advance preparation factors affecting results of sequence rule analysis in web log mining. E & M., **13**, 143-160 (2010).
[4] M. Munk and M. Drlík, Impact of different pre-processing tasks on effective identification of users’ behavioral patterns in web-based educational system. Procedia Computer Science. **4**, Elsevier, 1640-1649 (2011).
[5] M. Munk and M. Drlík, Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining. Communications in Computer and Information Science. Springer, **166**, 60-74 (2011).
[6] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD, **1**, 12-23 (2000).
[7] D. Pierrakos, G. Paliouras, Ch. Papatheodorou and C. D. Spyropoulos, Web usage mining as a tool for personalization: A survey. User Modeling and User-Adapted Interaction, Kluwer Academic Publishers. **13**, 311-372, (2003).

[8] F. Tao and F. Murtagh, Towards knowledge discovery from WWW log data, International Conference on Information Technology: Coding and Computing (ITCC'00) 302-307 (2000).

[9] Ch. Rana, A study of web usage mining research tools, Int. J. Advanced Networking and Applications, 3 1422-1429 (2012)

[10] M. Koutri, N. Avouris and S. Daskalaki, Chapter 7: A survey on web usage mining techniques for web-based adaptive hypermedia systems, in S. Y. Chen and G. D. Magoulas (ed), Adaptable and Adaptive Hypermedia Systems, IRM Press, Hershey, 125-149 (2005).

[11] R. Kosala and H. Blockeel, Web mining research: A survey. ACM SIGKDD (2000).

[12] F. M. Facca and P. L. Lanzi, Recent developments in web usage mining research, DaWak 140-150 (2003).

[13] F. M. Facca and P. L. Lanzi, Mining interesting knowledge from weblogs: a survey. Data & Knowledge Enginnering, 53, 225-241 (2005).

[14] R. Cooley, B. Mobasher and J. Srivastava, Data preparation for mining world wide web browsing patterns. Knowledge and Information System, **1** (1999).

[15] V. Chitraa and A. S. Davamani, A survey on preprocessing methods for web usage data, International Journal of Computer Science and Information Security, **7**, 78-83 (2010).

[16] T. T. Aye, Web log clearing for mining web usage patterns, 3rd International Conference on Computer Research and Development (ICCRD), 490-494 (2011).

[17] G. Castellano, A. M. Fanelli and M. A. Torsello, Log data preparation for mining web usage patterns. IADIS International Conference of Applied Computing 371-378 (2007).

[18] S. Tan, M. Chen and G. Yang, User behavior mining on large scale web data. International Conference on Apperceiving Computing and Intelligence Analysis (ICACIA), 60-63 (2010).

[19] M. A. Bayir, I. H. Toroslu and A. Cosar, A new approach for reactive web usage data processing, 22nd International Conference on Data Engineering Workshops (ICDEW 06) (2006).

[20] B. Mobasher, Chapter 15: Web usage mining and personalization, in M. P. Singh (ed), Practical handbook of internet computing, Chapman and Hall/CRC, **1444**, (2004).

[21] V. Chitraa and A. S. Thanamani, A novel technique for sessions identification in web usage mining preprocessing, International Journal of Computer Applications, **34**, 23-27 9 (2011).

[22] Z. Chen, A. W. Fu and F. Ch. Tong, Optimal algorithms for finding user access sessions from very large web logs, 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Springer-Verlag, London 290-296 (2002).

[23] C. Zhang and L. Zhuang, New path filling method on data preprocessing in web mining. Proceedings of Computer and Information Science 112-115 (2008).

[24] Y. Li, B. Feng and O. Mao, Research on path completion technique in web usage mining. International Symposium on Computer Science and Computational Technology, **1**, 554-559 (2008).

[25] L. Bing, Web data mining. Exploring hyperlinks, contents and usage data. Springer (2007).

[26] O. Nasraoui and E. Saka, Web usage mining in noisy and ambiguous environments: Exploring the role of concept hierarchies, compression, and robust user profiles, in B. Berendt et al. (eds), From web to social web: Discovering and deploying user and content profiles. Workshop on Web mining, WebMine 2006, LCNS **4737**, (2006).

[27] J. Song, T. Luo, S. Chen and F. Gao, The data preprocessing of behavior pattern discovering in collaboration environment. Int. Conference on Web Intelligence and Intelligent Agent Technology. 521-525 (2007).

[28] S. P. Nina, M. M. Rahaman, K. I. Bhuiyan and K. E. Ahmed, Pattern discovery of web usage mining, International Conference on Computer Technology and Development. 499-503 (2009).

[29] O. Nasraoui, M. Soliman, E. Saka, A. Badia and R. Germain, A web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE transactions on knowledge and data engineering. **20**, 202-215 (2008).

[30] H. Yilmaz and P. Senkul, Using ontology and sequence information for extracting behavior patterns from web navigation logs. IEEE International Conference on Data Mining Workshops, 549-556, (2010).

[31] O.R., Zaiane, Web Usage Mining for a better Web-based Learning Environment, Conference on Advanced Technology for Education, Alberta (2001).

[32] [O. R. Zaiane and J. Luo, Towards Evaluating Learners' Behavior in a Web-based Distance Learning Environment, IEEE International Conference on Advanced Learning Technologies (ICALT01), Madison, WI (2001).

[33] Ch. Nukoolkit, P. Chansripiboon and S. Sopitsirikul, Improving university e-learning with exploratory data analysis and web log mining. International Conference on Computer Science & Education (ICCSE 2011), Singapore, 176-179 (2011).

[34] F. Zhang, A New Method of Information Aggregation for Multi-type Evaluation Subjects-Involving Teachers' Performance Evaluation Based on TDW Operator Applied Mathematics and Information Sciences Special Issues, 901-906 (2012)

[35] Š. Koprda, P. Brečka and M. Milan, Project and realization of wifi net, Conference on Trends in Education Location. Olomouc, 301-304 (2009).

[36] Z. Balogh, M. Magdin and M. Turčáni, Interactivity elements implementation analysis in e-learning courses of professional informatics subjects, 8th International Conference on Efficiency and Responsibility in Education, Prague, 5-14 (2011).

[37] C. G. Marquardt, K. Becker and D. D. Ruiz, A pre-processing tool for web usage mining in the distance education domain. International Database Engineering and Applications Symposium (IDEAS '04). IEEE Computer Society, Washington, DC, USA, 78-87 (2004).

[38] F. B. Chanchary, I. Haque and M.S. Khalid, Web usage mining to evaluate the transfer of learning in a web-based learning environment, Workshop on Knowledge Discovery and Data Mining 249-253 (2008).

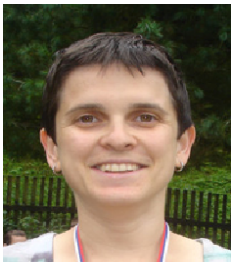
[39] V. Škorpil and J. Šťastný, Back-propagation and k-means algorithms comparison. 8th International Conference on Signal Processing. Guilin, China, IEEE Press. 1871-1874. (2006).

[40] Y. Yao, Credit risk assessment of online shops based on fuzzy consistent matrix. Applied Mathematics and Information Sciences **5**, 163S-169S (2011)
[41] N. Zhao, Y. Liu, Product approximate reasoning of online reviews applying to consumer affective and psychological motives research. Applied Mathematics and Information Sciences **5**, 45S-51S (2011)
[42] S. Weifeng, S. Mingyang, L. Xidong, L. Mingchu, An improved personalized filtering recommendation algorithm. Applied Mathematics and Information Sciences **5**, 69S-78S (2011)
[43] S. Iftikhar, F. Ahmad and K. Fatima, A Semantic Methodology for Customized Healthcare Information Provision. Information Sciences Letters **1**, 45-59 (2012)
[44] T.H. Tsai, H.T. Chang, Surfrom: A community-oriented search engine interface. Applied Mathematics and Information Sciences 6, 389-396 (2012)
[45] S. Araya, M. Silva and R. Weber, A methodology for web usage mining and its application to target group identification. J. Fuzzy Sets and Systems, **148**, 139-152, (2004).
[46] M. J. Berry and G. S. Linoff, Data mining techniques: for marketing, sales, and customer relationship management. John Willey & Sons (2004).
[47] I. Stankovičová, Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner. Informační a datová bezpečnost ve vazbě na strategické rozhodování ve znalostní společnosti 1-9 (2009).
[48] L. D. Catledge and J. E. Pitkow, Characterizing browsing strategies in the World-Wide Web. International World-Wide Web conference on Technology, tools and applications. 1065-1073 (1995).
[49] M. Munk, M., Vrábelová and J. Kapusta, Probability modeling of accesses to the web parts of portal. Procedia Computer Science. Elsevier, **3**, 677-683 (2011).
[50] M. Munk, M., Drlík and M. Vrábelová, Probability modeling of accesses to the course activities in the web-based educational system. Lecture Notes in Computer Science. Springer, **6786**, 485-499 (2011).
[51] J. Anděl, Základy matematické statistiky. MATFYZPRESS, Praha (2007).
[52] G. Rodríguez, Generalized linear models (2007).
[53] S. Elo-Dean and M. Viveros, Data mining the IBM official 1996 Olympics Web site. Technical report, IBM T.J. Watson Research Center (1997).
[54] M. S. Chen, J. S. Park and P. S. Yu, Data mining for path traversal patterns in a web environment. Proceedings of the 16th International Conference on Distributed Computing Systems 385-392 (1996).
[55] R. Cooley, B. Mobasher and J. Srivastava, Grouping web page references into transactions for mining world wide web browsing patterns. Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX '97). IEEE Computer Society, Washington (1997).
[56] M. Spiliopoulou, B. Mobasher, B. Berendt and M. Nakagawa, A framework for the evaluation of session reconstruction heuristics in web usage analysis. INFORMS Journal of Computing, **15**, 171-190 (2003).
[57] M. Munk, A. Pilková, M. Drlik, J. Kapusta and P. Švec, Verification of the fulfilment of the purposes of Basel II, Pillar 3 through application of the web mining methods. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. - 2012, 217-223 (2012).

Michal Munk received the PhD degree in Mathematics from the Department of Mathematics at Constantine the Philosopher University in Nitra, in 2007. He is currently an associate professor at the Computer Science Department of Constantine the Philosopher University in Nitra. His research interests are in the areas of web log mining and user behaviour modelling.

Martin Drlik received the PhD degree in Computer Science from the Department of Computer Science at Constantine the Philosopher University in Nitra, in 2009. He is currently an assistant professor at the Computer Science Department of Constantine the Philosopher University in Nitra. His research interests are in the areas of web log mining, educational data mining and learning analytics.

Jozef Kapusta received the PhD degree in Computer Science from the Department of Computer Science at Constantine the Philosopher University in Nitra, in 2009. He is currently an assistant professor at the Computer Science Department of Constantine the Philosopher University in Nitra. His research interests are in the areas of web log mining and user behaviour modelling.



Daša Munková received the PhD degree in Mathematics from the Department of Mathematics at Constantine the Philosopher University in Nitra, in 2007. She is currently an assistant professor at the Department of Translation Studies at Constantine the Philosopher University in Nitra. Her research interests are in the areas of text mining, linguistics and natural language processing.