IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 19, 2017, accepted May 5, 2017, date of publication May 23, 2017, date of current version June 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2706302

Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques

MICHAL MUNK[1], MARTIN DRLíK[1], L'UBOMíR BENKO[2], AND JAROSLAV REICHEL[1]
[1]Department of Informatics, Constantine the Philosopher University, SK-949 01 Nitra, Slovakia
[2]Institute of System Engineering and Informatics, University of Pardubice, CZ-532 10 Pardubice, Czech Republic

Corresponding author: Martin DrlíK (mdrlik@ukf.sk)

This work was supported by the Slovak Research and Development Agency under Contract APVV-14-0336.

ABSTRACT Educational data preprocessing from log files represents a time-consuming phase of the knowledge discovery process. It consists of data cleaning, user identification, session identification, and path completion phase. This paper attempts to identify phases, which are necessary in the case of preprocessing of educational data for further application of learning analytics methods. Since the sequential patterns analysis is considered suitable for estimating of discovered knowledge, this paper tries answering the question, which of these preprocessing phases has a significant impact on discovered knowledge in general, as well as in the meaning of quality and quantity of found sequence patterns. Therefore, several data preprocessing techniques for session identification and path completion were applied to prepare log files with different levels of data preprocessing. The results showed that the session identification technique using the reference length, calculated from the sitemap, had a significant impact on the quality of extracted sequence rules. The path completion technique had a significant impact only on the quantity of extracted sequence rules. The found results together with the results of the previous systematic research in educational data preprocessing can improve the automation of the educational data preprocessing phase as well as it can contribute to the development of learning analytics tools suitable for different groups of stakeholders engaged in the educational data mining research activities.

INDEX TERMS Computational and artificial intelligence, data preprocessing, educational technology, learning, learning systems, sequential analysis, web mining.

I. INTRODUCTION

Copmuter-Based Education (CBE) means using computers in education to provide guidance, to instruct or to manage instructions to the student [1]. Today, the CBE often takes the form of web-based educational systems, which together with artificial intelligence techniques, have induced the emergence of new educational systems such as learning management systems (LMS), intelligent tutoring systems (ITS), adaptive hypermedia educational systems (AHES), personal learning environments (PLE), massive open online courses (MOOC), etc. These systems have many common features from the educational data storing and preprocessing point of view. Therefore they will be denoted as virtual learning environments (VLE) for the purpose of this paper.

VLEs are deployed in all sectors of education. They gather a vast amount of different tracking data that have to be preprocessed in different ways depending on both the nature of available data and the specific problems and tasks to be resolved by data mining techniques [1]. However, although virtual learning environments gather stakeholders' data automatically, exploitation of the data for learning and teaching is still insufficient [2].

Many researchers have applied data mining methods and techniques on educational data in the last few years considering the success of an application of data mining methods in other domains. They aimed to help all stakeholders of the VLEs to improve their teaching and learning competences, support managerial tasks, as well as design, create and develop more efficient and attractive e-learning courses. This trend gave the basis to the new research disciplines Learning Analytics (LA) [3] and Educational Data Mining (EDM) [4].

EDM and LA have many common features. A typical EDM, as well as LA process, converts raw data coming from VLEs into useful information that could potentially have a great impact on educational research and practice [5]. They both start with the data gathering and preprocessing steps, which have many similarities. On the other hand, Siemens and Baker [6] identified five key areas of difference between the EDM and LA, including a preference for automated paradigms of data analysis (EDM) versus making human judgment central (LA), a reductionist focus (EDM) versus a holistic focus (LA), and a comparatively greater focus on automated adaptation (EDM) versus supporting human intervention (LA) [7]. However, these differences do not affect the initial phases of educational data analysis too much. Therefore, the outcomes of this paper, which is primarily focused on the educational data preprocessing in VLEs, could be considered applicable as well as useful for both research disciplines.

In data gathering step, data is collected from different stakeholders' activities when they interact with the learning objects within the VLEs [8]. Gathered data often takes the form of logs (records) stored in the relational database tables or text files with conventional format. A log file can be thought of as a list of a VLE stakeholder's events, in which each line or record contains a time-stamp plus one or more fields that hold information about activity at that instant.

Relevant information can be extracted through the analysis of these logs that may help in understanding many educational and managerial processes within VLEs. Lara et al. pointed out that activity logs stored in VLE have proved to be a useful source of data for data mining. In the particular case of VLE Moodle, which will be used in the experiment described in this paper, several earlier studies indicate that this is an environment in which data mining techniques have proved to be useful [9]. Lavigne *et al.* [10] came to the similar conclusion. However, they stated that the analysis has been effective for questioning certain aspects of online learning, but in some cases, the implementation of the findings into the VLEs does not yield the expected results.

Data preprocessing is the first step of the EDM or LA analytical process. It transforms raw data from the logs into a shape suitable for resolving a problem using a specific data mining method, technique or algorithm. This step is often considered the most time-consuming step, which requires a significant effort and consumes the greater part of the available resources [11].

Several terms should be defined because of their frequent use in the paper. According to Chitraa et al., data preprocessing step consists of four separate phases: data cleaning, user identification, session identification and path completion [12]. A session can be defined as a semi-permanent interactive information interchange between two or more communicating devices, for example, a login session is the period of activity between a user logging in and logging out.

A user session is defined as a sequence of requests made by a single user over a certain navigation period, and a user may have a single (or multiple) session(s) during this time period. According to Spiliopoulou *et al.* [13], a user session can be defined as a sequence of necessary steps required to fulfil a particular task successfully. Session identification is a process of segmenting the log data of each user into individual access sessions. In other words, the session is a sequence of VLE user's steps leading to a particular learning aim. Although user and session identification is not specific to education, it is especially relevant due to the longitudinal nature of student usage data [1].

A path completion means a reconstruction of missing activities of a VLE's stakeholder. Reconstruction of these activities is focused on retrograde completion of records (logs) on the navigation path went by the stakeholders in the VLE. The main aim of this step is adding missing records to the input data file, which are not automatically recorded in the log file or database records.

The paper is focused on the common tasks and problems related to the session identification and path completion phases of data preprocessing of the EDM and LA processes. The main objective is to evaluate the impact of the different settings used in the preprocessing phase of LA and EDM on the quality of the discovered knowledge. Specifically, the main aim is to assess the impact of user session identification and path completion on the quantity and quality of extracted sequence rules that represent the learners' behavioral patterns in the VLE. Moreover, it tries to answer the question whether and how could be the preprocessing phase accelerated and at least partially automated using special tools as well as tools integrated into the VLEs.

The rest of the paper is structured as follows. The next section summarizes the actual state of the research in educational data preprocessing from several points of view. It tries to justify the fact that the lack of attention is paid to the particular steps of educational data preprocessing from log files, although these steps form an inevitable part of the EDM research and specialized educational data preprocessing tools. The third section provides a comprehensive description of the research methodology and articulates the research assumptions. It begins with the summarizing of the results of previous experiments focused on the evaluation of the importance of different preprocessing phases on the quality of the discovered knowledge. These results complement the results of the described experiment, which are introduced in the fourth section. These results provide the reader a solid knowledge, how to evaluate the contribution of different preprocessing techniques widely used in the educational data preprocessing phase to the overall knowledge discovery. The last section provides the discussion about the contribution and weaknesses of the experiment and suggests several ideas for the future research.

II. RELATED WORK

Scientific progress in EDM research area can be found in reviews [4], [14]. The last comprehensive state-of-the-art review was edited by [1] and [15]. Scientific progress in the

LA research discipline is summarized in [16]–[18]. They are focused mainly on the application, comparison as well as improvement of different data mining methods, techniques, and tools used usually in the later phases of the EDM as well as LA processes. However, the preprocessing problems are mentioned only partially. For instance, the last book about the progress in the LA edited by Larusson and White [19] does not mention educational data preprocessing problems at all. The authors mostly assumed, there are only small problems, which can follow the preprocessing phase.

It can be stated, that educational data preprocessing has not received sufficient analytical efforts based on the analysis of these comprehensive EDM and LA reviews. According to [17], data preprocessing is the first step in any data mining process, being one of the most important but less studied tasks in educational data mining research. Most of the authors rarely describe this important step or only provide a list of few works focused on the preprocessing of data. This fact can be considered surprising because according to several authors, data preprocessing phase typically requires a significant effort and can take typically between 60 and 90 percent of the resources of the knowledge discovery process [11]. Similarly, Bienkowski et al. supposed that at least 70 percent and often 80 to 85 percent of the effort in data analytics is devoted to data cleaning, formatting, and alignment. Moreover, they suggested that education has the further complication of needing to move data across different levels of the system, back and forth between classroom, school, district, and state databases [20].

Special attention to the preprocessing phase of educational data analysis is paid mainly in [21]. Romero *et al.* [5] wrote a special chapter there focused especially on data preprocessing. They summarized specific problems about data preparation in web-based educational systems and provided references to other relevant scientific papers. They noticed most of the studies aiming to apply data mining techniques in LA and EDM are based on the preprocessing technique employed in e-commerce. However, the context of e-learning is very particular and differs from ordinary websites or e-commerce contexts, at the level of the structure, in the nature of the contents, or in the objectives of analysis. A specific characteristic of educational data is that there are different levels of granularity such as keystroke level, answer level, session level, student level, classroom level, and school level. They also stated, similarly as already mentioned authors, that there is still a lack of specific references to the educational data preprocessing in the current scientific literature. The authors finally intervened for enhancing preprocessing facilities that prepare the e-learning data in a meaningful and useful way [5].

The same authors stated, the number of tools exclusively devoted to data preprocessing is marginal [5]. Nowadays, general software and data mining tools are used for educational data preprocessing. However, most of the current EDM and LA tools and general preprocessing tools are normally designed more for power and flexibility than for simplicity. They do not suitably support preprocessing activities in the educational domain. Moreover, most of the currently existing tools are just prototypes providing restricted features, or they are oriented to work only with a very specific type of data.

Nevertheless, some prototypical proposals for data preprocessing solutions can be found in the field of education. These tools are mainly oriented to the preparation of data extracted from log files in VLE. Usable integrated tools for teachers that support cyclical research activities are still missing in most current VLEs. Even though they exist, they are far from satisfactory [8]. These tools support predominantly the modelling and visualization phase of the knowledge discovery [22]. Moreover, they require more experienced users, who are able to configure data mining algorithms before they are executed [23]. Specialized preprocessing tools, which could support or automate all mentioned preprocessing tasks, are still missing and must be done manually.

Marquardt *et al.* [24] published a comprehensive paper about the preprocessing of educational data. They defined a learning session, which can span different periods according to given circumstances or learning goals. They proposed a tool prototype for the automation of the most typical tasks performed in the preprocessing phase for the mining of educational data like data cleaning and filtering, user and session identification, path completion, data enrichment, and transaction identification [25]. The tool prototype is not currently widely used in connection with EDM research.

The educational data used in this paper come from the VLE Moodle. VLE Moodle belongs to the mostly used VLEs for several years. Therefore, it is not surprising, that many researchers focused their research on the implementation of data mining and especially web mining methods onto educational data recorded in this system [26]. VLE Moodle is used as a coherent framework of preprocessing, and data extracted from VLE Moodle have served as a case under study in most examples [5].

VLE Moodle does not offer an integrated tool for data analysis using LA or EDM techniques and methods. It only provides access to the students' logs at the e-learning course level as well as at the system level. The official Moodle site contains a comprehensive list of learning analytics plugins (SmartKlass, Zoola, Engagement analytics, GISMO, Configurable reports), which can be installed into or connected with the Moodle. However, these plugins can be considered more reporting tools than analytical tools in the meaning of LA and EDM tools.

Dimopoulos *et al.* [27] summarized the EDM and LA tools which interoperate with Moodle. However, these tools (for instance, LAe-R, MOCLog) provide mainly analysis and visualization of the educational data and combine a didactical theory with VLE stakeholders' requirements [28]. They do not deal with the educational data preprocessing in detail. The author did not mention which steps of the data preprocessing in VLE Moodle are necessary.

Ali *et al.* [22] introduced semantic learning analytics tool named LOCO-Analyst, which provide educators with feedback on the relevant aspects of the learning process taking place in a web-based learning environment. They similarly do not deal with the preprocessing phase in detail.

As was stated previously, data preprocessing step consists of data cleaning, user identification, session identification and path completion phases. Data cleaning is often minimal because most VLEs provide user authentication and logging his/her activities in a suitable form. Users are identified mainly by login and password. They have a unique user ID. Romero et al. stated that therefore it is not necessary to do the typical user identification task to identify sessions from logs, and session determination ceases to be a problem. They argued that all records can be sorted in ascending order with the user ID as the primary key, and the event time as a secondary key. As a result, it is easy to identify user sessions by grouping contiguous records from one login record to the next one. Moreover, an upper limit of the time interval between two successive clicks denoted as session timeout threshold (STT) [13], can be set in order to break the sequence of one student's click stream into sessions. Consequently, the changes in the value of STT may result in increasing or decreasing the total number of identified sessions. However, as they stated, there is no research on the relation between this time limit of user session identification and its impact on quality of discovered knowledge [5].

This paper is focused on the user session identification and path completion techniques used in the preprocessing step. A user session can be defined as a set of pages visited by the same stakeholder within the duration of one particular visit to a VLE. Once a stakeholder was identified, his/her click stream is portioned into logical clusters. The method of portioning into sessions is called in general sessionization or session reconstruction [12]. Session reconstruction techniques can be divided into time-oriented heuristic techniques and navigation oriented heuristics techniques. These techniques use reactive strategy because they exploit background knowledge on user navigational behavior to assess whether the records registered in the log file can belong to the same individual and whether these records were performed during the same or subsequent visits of the individual to the VLE [13]. The selection of the particular heuristic depends on the design of the VLE as well as on the assumed average duration of the stakeholders' session.

Even though different techniques of session identification and path completion are extensively described in the web usage mining reviews [12], [29], [30] as well as in several reviews of the EDM research field [14], [21], they focused mainly on improving of existing techniques, or application of these techniques in specific environments. They did not deal with the impact of particular preprocessing phases and techniques on quality and quantity of discovered knowledge in general as well as regarding sequential patterns analysis [5].

Sequential patterns analysis is a process of discovering and displaying previously unknown knowledge, interrelationships, and data patterns with the goal of supporting improved decision making. Sequential patterns analysis can contribute to the educational research in various ways. It helps to evaluate learner's activities, adapt and customize resources, compare theoretical and real learning paths, generate personalized activities to different groups of learners or recommend to a student the most appropriate educational material [31]. Therefore, the sequential patterns analysis, especially found sequence rules, is considered suitable for estimating of discovered knowledge [32].

Sequence rules are defined as a consecutive or non-consecutive ordered sub-set of an events sequence. Although the mining of the complete set of sequence rules has been improved substantially, in many cases, the sequential pattern still faces tough challenges in both effectiveness and efficiency. The problem is that there could be a large number of sequential patterns in a large database. However, presenting the complete set of sequential patterns may make the mining result hard to understand and hard to use. A user is often interested in only a small subset of such patterns [33].

According to Berry and Linoff [34], extracted sequence rules can be divided into three groups: actionable (useful), trivial and inexplicable rules. Useful rules contain high quality and actionable information. Trivial rules are known to anyone who is familiar with the business or domain. Inexplicable rules seem to have no explanation and do not suggest any action.

The quality of found sequence rules is possible to assess using two indicators [34]: variables *support* and *confidence*. *Support* of a rule is defined by a fraction of transactions that satisfy union of items in the consequent and antecedent of the rule [35]. The value of *support* for a rule "if A then B", where A, B are itemsets (sequences), can be calculated as:

$$\begin{aligned} &support\ (if\ A\ then\ B) \\ &\quad = \frac{frequency\ of\ (A,\ B)}{number\ of\ transactions\ in\ the\ dataset} * 100. \end{aligned}$$

Support corresponds to statistical significance and on the other hand *confidence* is a measure of the rule's strength [35], where *confidence* can be defined as:

$$confidence\ (if\ A\ then\ B) = \frac{support\ (if\ A\ then\ B)}{support\ (A)} * 100.$$

The *confidence* for rule "if A then B" is not necessarily the same as the *confidence* of the rule "if B then A" [36].

It can be stated considering the provided review of the related work, that only several papers, including the previous research of the authors, systematically study the impact of different parameters of the preprocessing techniques on the quality of discovered knowledge [37]. The results of the previous research will be explained as the background of the research in the next section because they naturally complement to the described experiment.

III. RESEARCH METHODOLOGY

This section describes the experiments, in which different session identification and path completion data preprocessing techniques were applied with the aim to find out, which of them are necessary to realize during the EDM or LA process.

A. RESEARCH BACKGROUND

Two previous experiments [38], [39], which closely relate to the aim of this paper, were focused on the specification of inevitable steps of data preprocessing with the aim to obtain reliable data from the log files of the VLE. These experiments, together with the experiment described later, provide the reader with a solid review of the problems related to the selection of suitable preprocessing techniques in the context of EDM as well as LA. The summarized results could help researchers to decide, which steps, techniques, as well as their parameters, use in the educational data preprocessing, especially in the case of VLE Moodle.

The first experiment was focused on an evaluation of the impact of different variables (user ID, IP address, time) suitable for session identification on knowledge discovery represented by the behavior patterns of students in the VLE [38]. Surprisingly, the results showed that paths completion in combination with all three variables (user ID, IP address, time) had no significant impact on the quantity and also on the quality of obtained knowledge. Completing the paths had the impact only on increasing the portion of useful rules, but this increase was not statistically significant.

On the other hand, session identification based on the variable *time,* often denoted as *session timeout threshold* (STT), had a significant impact on the quantity as well as on the quality of obtained knowledge. According to Spiliopoulou *et al.* [13], this technique belongs to the time oriented heuristics because it uses an upper bound on time spent in the entire VLE during a visit. The value of STT is often used to determine when a session ends and the next one starts. An STT is a pre-defined period of inactivity that allows web applications to determine when a new session occurs [40]. The correct value of*STT* has been often discussed by several authors in the web mining research field [41] as well as EDM [5]. They recommended an interval of *STT* values found using different criteria or statistical estimations. However, no generalized model was proposed as well as proved to estimate the *STT* used to identify sessions in VLEs.

The portion of trivial and inexplicable rules was dependent on session identification based on time, e.g. on the value of *STT*. The identification of sessions had a significant impact on the reduction of a portion of trivial and inexplicable rules while the portion of useful rules stayed unchanged.

The second experiment [39] was therefore focused on the evaluation of the impact of the session identification techniques based on the different session timeout threshold variable on the quantity and quality of discovered knowledge. The VLE stakeholders' logs were exported and modified with the aim creating several log files with various values of *STT* and path completion.

The assumption concerning the identification of sessions based on time and its impact on the quantity of extracted rules was proved. Moreover, it was proved that the value of *STT* has a significant impact on the quantity of extracted rules. Statistically significant differences in the average of variables *incidence*, *support*, as well as the *confidence* of found rules were proved among files with different *STT* regardless of the fact, whether the files were modified by path completion. The portion of trivial and inexplicable rules was dependent on the value of *STT*. Identification of sessions based on smaller *STT* had an impact on reducing the portion of trivial and inexplicable rules as well as on the quality of found rules in terms of the fundamental characteristics of quality [37]. On the other hand, it was shown that the completion of paths had a neither significant impact on the quantity nor quality of extracted rules. Paths completion had no significant impact on increasing the portion of useful rules.

As a result, path completion in connection with the impropriate value of *STT* identification may cause increasing of trivial and inexplicable rules. The results showed the highest degree of concordance in the variables *support* and *confidence* of the sequence rules found in the file without path completion and in the corresponding file with the path completion. The assumption of an impact of paths completion on obtained knowledge was not proved. It can be stated considering the previous results, that the identification of sessions based on time is crucial to data preparation from a log file in the VLE [42]. However, the correct estimation of the variable *length* of *STT* upon identifying sessions based on time is also important. The decision to use too high value of *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 [37], [38].

B. DATA ACQUISITION

Eight log files in different stages of data preprocessing were compared in the following experiment. The raw log files came from an e-learning course, which was created in VLE Moodle for the purpose of a blended learning methodology support of the course Computer Data Analysis (CDA). Eighty students enrolled in the e-learning course. The primary course objective was to familiarize students with different study programs with the introductory topics of data mining, inference, and exploratory analysis [43]. The didactical objective of the e-learning course was to verify whether and how this subject, accompanied by an e-learning course with practical assignments solved in the given statistical software, can increase the statistical literacy of the students [44].

While the structures of the VLE Moodle logs were changed [28], [45], it was necessary to modify the methodology used in the experiments mentioned in the previous section and also to identify the variables needed for the next experiment (*IP address*, *date*, and *time* of access, *user ID*, *URL*). Other variables were removed from the log files. An example of the structure of the log file is depicted in Table 1.

The *URL* variable did not involve the web domain of the VLE and also was not in a typical HTTP format. It gave only a reference if the user visited a particular course activity or resource, like a book, an exam, a dictionary or if

TABLE 1. The logs file from e-learning course.

URL	Action	Type	URL ID	IP Address	User ID	Unix Time
\mod_forum\event\discussion_viewed	viewed	discussion	39103	85.162.202.190	2347	1415566779
\core\event\course_viewed	viewed	Course	661	85.162.202.190	2347	1415566983
\mod_book\event\chapter_viewed	viewed	chapter	63270	92.245.193.165	4855	1415631993
\mod_quiz\event\attempt_started	Started	attempt	64853	10.160.0.106	5213	1418635040
\mod_quiz\event\attempt_viewed	viewed	attempt	64853	10.160.0.106	5123	1418635050
\mod_quiz\event\attempt_submited	submitted	attempt	64853	10.160.0.106	5123	1418635379
⋮	⋮	⋮	⋮	⋮	⋮	⋮

TABLE 2. The logs file from e-learning course.

URL ID	Referrer
69476	661
63270&chaptered=29658	63270
39101	661
70323	661
⋮	⋮

he/she posted an assignment. The structure of the URL is described in detail in the columns in Table 1. For better clarity of the visited URL, the variable was shortened only to its corresponding ID and was created as a new variable URL ID. This variable was further used in the next data preparation steps.

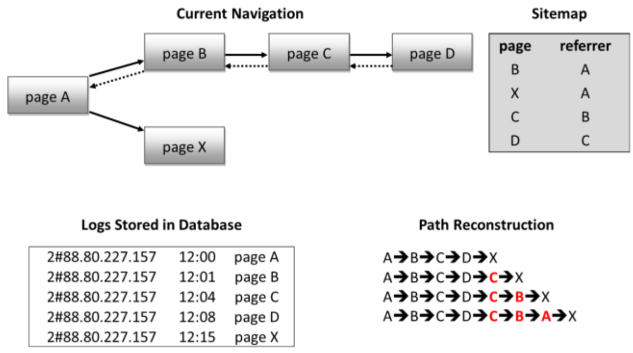
C. CREATION OF DATA MATRICES

The data matrices were created from the log files, which contained the information about the students' accesses ordered by the *user ID*, *IP address*, and *time* variable.

Consequently, a sitemap of e-learning course was also created. This sitemap was used in connection with several session reconstruction techniques later in the experiment. The sitemap collects information about the structure of the e-learning course content and navigation. It has a great importance for retrograde completion of the records on the path went through by the user using a back button since the use of such button is not automatically recorded into log entries of VLE. The information on the existence of links to particular pages of the e-learning course can be extracted from the sitemap.

The sitemap was obtained using Web Crawling application implemented in the used STATISTICA Data Miner. The sitemap had to be modified (Table 2) to correspond with the IDs of e-learning course pages. Having ordered records according to the IP address it was possible to search for some linkages between the consecutive pages.

For example, a sequence of pages for the selected IP address can look like this: A→B→C→D→X (Fig.1). The algorithm can find out if there exists the hyperlink from the page D to the page X using the sitemap. The assumption is that page X was accessed by the same user using a back button of the web browser from one of the previous pages. Then, through a backward browsing can be found out, from which of the previous pages exists a reference to page X. Fig. 1 shows, that there is no 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, it is possible to find out, that there is not any hyperlink connection between page B and page X and so B can be added into the sequence, i.e. A→B→C→D→C→B→X. Finally, the algorithm finds out that page A contains a hyperlink to page X. After the termination of the backward path analysis, the sequence will look like this: A→B→C→D→C→B→A→X. It means the user used the back button in the browser in order to cross from page D to C, from C to B and from B to A [39].



page	referrer
B	A
X	A
C	B
D	C

2#88.80.227.157	12:00	page A
2#88.80.227.157	12:01	page B
2#88.80.227.157	12:04	page C
2#88.80.227.157	12:08	page D
2#88.80.227.157	12:15	page X

FIGURE 1. Example of path reconstruction.

The variable *STT* was created in the log file based on a time window of 100 minutes as a next step in the data matrices creating process. This value was chosen regarding the duration of an average face-to-face lesson. The lesson normally took 90 minutes, but sometimes the students finished their tasks after the lesson, during the break, therefore more ten minutes were added to the value of *STT* estimation as a period of latency. It was proved, that the selected additional ten minutes were after rounding in accordance with the value of a standard deviation.

D. DATA PREPARATION

As was stated earlier, all preprocessing steps were not necessary to apply in the presented experiment because the VLE did not allow an anonymous login into e-learning courses. All users were already identified by user ID. Teachers' and course administrator's records in log file were removed. Only the activity of the students enrolled in the e-learning course was recorded in the log files.

The experiments mentioned in section 3.1 used time-oriented heuristic techniques of user session reconstruction. On the other hand, the navigation oriented heuristic using reference length was used in the next experiment. The Reference Length method is based on the assumption that the time the user spends on a page correlates with whether the page is classified as an auxiliary (navigational) page or content page for that user [46], [47]. It can be assumed based on the research in [47] that the time spent on the auxiliary pages is small. It is not expected, the auxiliary page of e-learning course contains any relevant educational content. The user passes through these pages so that he/she can navigate to his/her search target content. If the portion of auxiliary pages is known, then the maximum length of the auxiliary page is given by the formula $C = -\ln(1-p)/\lambda$, where C denotes cutoff time, i.e. a threshold value of auxiliary pages, p represents the percentage of auxiliary pages. Formula $\hat{\lambda} = 1/R\bar{Length}$ represents the maximum likelihood estimation of the parameter λ, where $R\bar{Length}$ is observed mean time spent on the pages [46].

If the cutoff time has been estimated, then the session is a sequence k of visited pages, where the first $k-1$ pages are classified as auxiliary pages because the time spent on these pages is less or equal to the cutoff time. The last page in this sequence is classified as a content page because the time spent on this page is higher than the cutoff time. The influence of the ratio of auxiliary pages on the cutoff time calculation can be compared based on the sitemap (objective estimate) and subjective estimate.

The decision that the page is auxiliary is subjective. It is based on what the e-learning course developer or teacher defines as an auxiliary page.

This estimation can also be used when the sitemap of the e-learning course is missing. On the other hand, when the ratio of auxiliary and content pages from a sitemap is calculated, it gave more accurate information about the ratio of auxiliary pages. The calculation was made using the formula $p = a/n$, where a was a count of auxiliary pages, and n represented a count of all pages.

The sitemap was created using a crawling tool and consisting of the variables *URL* and *Referrer*. The sitemap was imported into a database. The number of auxiliary pages from the variable *Referrer* was obtained using the SQL distinct statement. The number of auxiliary pages corresponded to the number of unique referring pages in the used VLE.

Finally, applying the previously described steps and session identification techniques led to the creating of four preprocessed log files (Fig.2):

- File A1 contained the sessions identified using the Reference Length method, and the ratio of auxiliary pages was calculated from the sitemap (15.32 %).
- File B1 contained the sessions identified using the Reference Length, and the ratio of auxiliary pages was subjectively estimated (25 %).
- The next session identification technique used in the experiment is based on an average length of a session timeout threshold. A new session was defined when the time between two actions was higher than the average length. The File C1 was created in this way.
- The last File D1 was created by the application of the method, in which the sessions were identified using the quartiles. The formula $Q_{STT} = Q_{III} + 1.5Q$ was used, where Q_{III} represented an upper quartile and Q was a quartile range. A new session was defined when the time between two actions in the e-learning course was higher than Q_{STT}.

The next step of data preparation was focused on the preparation of log files using path completion technique. It was necessary to identify important page accesses that were not recorded in the log file due to a client-side caching of e-learning course pages. This happened for instance when the user accessed to the page using a back button of the browser.

The sitemap of the VLE was created during the process of creating data matrices. It is necessary to know the sitemap of the VLE and the variable *Referrer* from the log file for adding the missing references to the log files. If the referrer URL is not the same as a previously requested page of the e-learning course, then the path is incomplete. Concerning problems with the variable *URL* of the log file, it was also necessary to modify the sitemap *URL* and *Referrer* variables. Both variables were shortened only in the case they corresponded to the e-learning course main page, book, exam, dictionary or assignment. Details of the used path completion technique were described in section 3.3 as well as in [30] and [47]. Based on them, the path completion algorithm was applied in the last step of data preparation. Finally, another four files were created for every previously used session identification techniques (File A2, File B2, File C2, and File D2).

E. DATA ANALYSIS

The main goal of data analysis stage of this experiment represents a searching for behavioral patterns of students in individual log files. A sequence patterns analysis was applied to all prepared log files with the aim to extract sequence rules for each file. As a result, a set of extracted sequence rules from the frequent sequences with the minimum of variable *support* 1 % for each file was created. As was mentioned earlier,

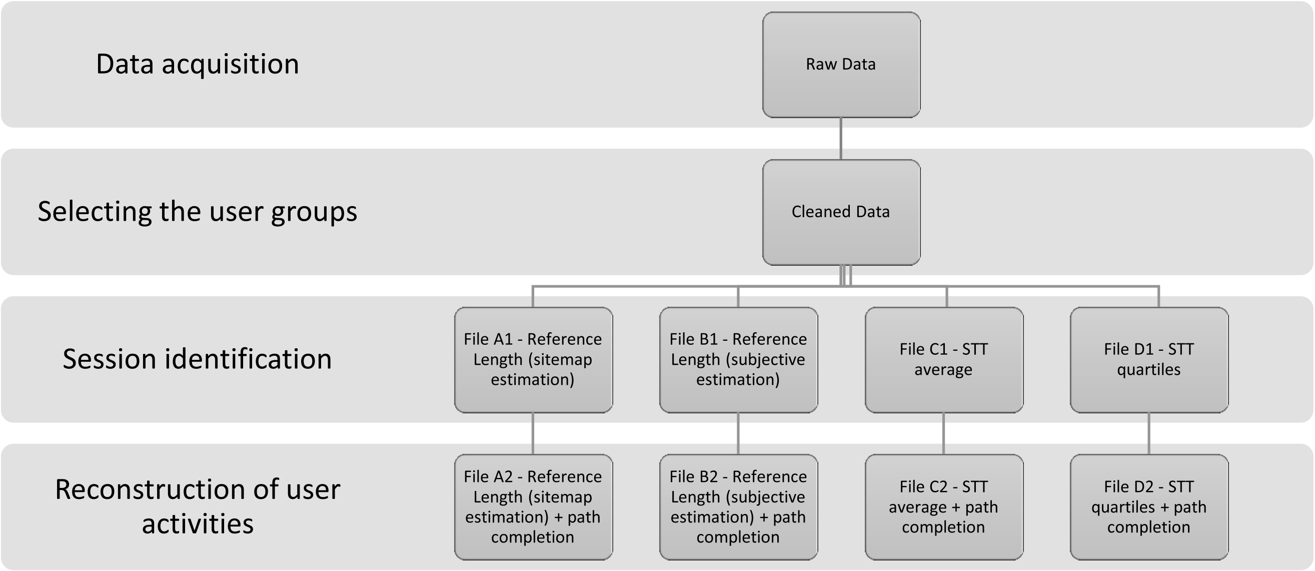

Data acquisition
Raw Data
Selecting the user groups
Cleaned Data
Session identification
File A1 - Reference Length (sitemap estimation)
File B1 - Reference Length (subjective estimation)
File C1 - STT average
File D1 - STT quartiles
Reconstruction of user activities
File A2 - Reference Length (sitemap estimation) + path completion
File B2 - Reference Length (subjective estimation) + path completion
File C2 - STT average + path completion
File D2 - STT quartiles + path completion

FIGURE 2. Application of data preparation to the log file.

extracted sequence rules were divided into three groups: actionable (useful), trivial and inexplicable rules.

In this paper, trivial rules were mostly rules that included, for example, a transition from the main page to a book page, quiz page or assignment in the e-learning course and vice versa (e.g. Course => Autotest3). Inexplicable rules were identified as the transition from one page to the same one (e.g. Course => Course). Useful rules were identified as the transition from the final exam quiz page to the book page and consequently back to the final exam quiz page because it is not allowed during the final exam (e.g. Topic3/Autotest2, Topic2/Task2, Activity of lesson => Topic3/Autotest2).

The decision, to which type of sequence rules the given rule belongs, was solely subjective based on the domain expert's decision. Considering the subjective character of the expert's decision, its objectivity had to be evaluated. The objectivity denotes the measure to which the results are independent on the researcher as well as on the measured unit in the meaning of the distortion of the measurement. As a consequence, eighty nine found sequence rules evaluated by the eight human experts were selected. Moreover, the experts were divided into three groups based on their roles in the VLE and particularly in the e-learning course: four teachers (T), two e-learning course developers (C) and two VLE managers (M). All participants evaluated the usefulness of found sequence rules using a three-value scale. The overall evaluation of the rule was calculated as a weighted mean of the individual evaluations. This approach took into account the uneven distribution of the groups of experts.

The non-parametric method of analysis was used considering the unknown distribution of data and ordinal characteristic of the used variables. Finally, Kendall's tau was used. Statistical significance of the calculated coefficients was tested (Table 3).

Table 3 shows that there is a large almost ideal linear dependency (0.7 – 1) between the particular expert's evaluations of the sequence rules.

The coefficients of Kendall's tau were statistically significant ($p < 0.05$) between the particular expert groups (Table 4). In other words, it can be stated that the high coefficients of the correlation ensured the objectivity of the sequence rules evaluation. The highest rate of compliance was between the e-learning course developers (C) and VLE managers (M). On the other hand, the smallest rate of compliance was among the teachers (T) and e-learning course developers. The zero hypothesis supposed that there was not statistically significant difference in sequence rules evaluations between the groups of experts. This hypothesis was not rejected considering the results of the Friedman test (ANOVA Chi Sqr. ($N = 89$, $df = 2$) $= 2.294118$; $p = 0.31757$).

F. OUTPUT OF DATA UNDERSTANDING AND RESULTS COMPARISON

This step is characterized by creating of data matrices from the analysis and by defining the assumptions. The following assumptions were articulated:

- It is expected that an identification of sessions using the Reference Length method, calculated from a sitemap, will have a significant impact on the quantity of extracted rules.
- It is expected that an identification of sessions using the Reference Length method, calculated from a sitemap, will have a significant impact on an increasing the portion of useful rules.
- It is expected that an identification of sessions using the Reference Length method, calculated from a sitemap, will have a significant impact on the quality of extracted rules.

TABLE 3. Kendall's coefficients tau between the particular expert's evaluations of the sequence rules.

	T1	T2	T3	T4	C1	C2	M1	M2
T1	1.000	0.772	0.843	1.000	1.000	0.912	1.000	0.978
T2	0.772	1.000	0.723	0.772	0.772	0.739	0.772	0.757
T3	0.843	0.723	1.000	0.843	0.843	0.734	0.843	0.824
T4	1.000	0.772	0.843	1.000	1.000	0.912	1.000	0.978
C1	1.000	0.772	0.843	1.000	1.000	0.912	1.000	0.978
C2	0.912	0.739	0.734	0.912	0.912	1.000	0.912	0.892
M1	1.000	0.772	0.843	1.000	1.000	0.912	1.000	0.978
M2	0.978	0.757	0.824	0.978	0.978	0.892	0.978	1.000

TABLE 4. Kendall's coefficients tau between the particular expert groups' evaluations of the sequence rules.

	Valid N	Kendall Tau	Z	p-value
C & M	89	0.919	12.7536	0.0000
T & M	89	0.795	11.0269	0.0000
T & C	89	0.741	10.2785	0.0000

- It is expected that a path completion will have a significant impact on the quantity of extracted rules.
- It is expected that a path completion will have a significant impact on an increasing the portion of useful rules.
- It is expected that a path completion will have a significant impact on the quality of extracted rules.

Comparison of data analysis results elaborated on various levels of data preprocessing in terms of quantity and quality of the found rules will be closely described in the next section.

IV. RESULTS

This section summarizes the results of the experiment. The first part of the section deals with the evaluation of the quantity of extracted sequence rules in the files. The second part is focused on the qualitative evaluation of the extracted sequence rules.

A. QUANTITY EVALUATION OF EXTRACTED RULES IN THE EXAMINED FILES

The analysis (Table 5) resulted in sequence rules, which were obtained from frequent sequences fulfilling their minimum of variable *support* (in our case. min $s = 1$ %). Frequent sequences were obtained from identified sequences, i.e. visits of individual users during one term (Table 5). The STATISTICA Sequence, Association and Link Analysis module was used for sequence rules extraction. It is an implementation of a powerful A-priori algorithm used in several other experiments [35], [48]–[50] together with a tree structured procedure requiring one pass through the data.

There is a high coincidence between the results (Table 6) of sequence patterns analysis in terms of the portion of found rules in the case of files without path completion (A1, B1, C1, D1) and files with path completion (A2, B2, C2, D2), where 1 means the rule was found in the examined file and vice versa, 0 means the rule was not found in the examined file. The most rules were extracted from files with path completion; specifically, 76 rules were extracted from the file C2 which represents over 85 %. 72 rules were extracted from the file A2 which represents almost 81 %. 67 rules were extracted from the file B2 which represents over 75 %, and 58 rules were extracted from the file D2, which represents over 65 % of the total number of found rules. Generally, more rules were found in the observed files with the completion of paths. Considering the fact, the files had been created from the same set of logs, it is natural, that the found sequence rules overlapped partially.

Based on the Q test results of the zero hypothesis, which reasons that the incidence of rules does not depend on individual levels of data preparation, is rejected at the 0.1 % significance level (Table 6).

Kendall's coefficient of concordance represents the degree of concordance in the number of found rules among examined files. The value of the coefficient (Table 7) is approximately 0.26 while 1 means a perfect concordance and 0 represents a discordance. Low values of the coefficient confirm the Q test results.

Four homogenous groups (Table 7) from multiple comparisons (Tukey test) were identified regarding the average incidence of found rules. Statistically significant differences were proved at the 5 % significance level in the average incidence of found rules between files D1 and C1, D2 and C2 as well as between files without (X1) and with (X2, where X = {A, B, C, D}) path completion.

The path completion has a significant impact on the quantity of extracted rules. On the contrary, the session identification technique using reference length based on sitemap estimation has no impact on the quantity of extracted rules in the case of files without as well as with path completion.

A closer look at the results (Table 8, Table 9, Table 10, Table 11) shows that

- almost 50 % of the new rules were found in files with path completion (A2, B2, C2, D2),

TABLE 5. Number of accesses, sequences and rules.

	A1	B1	C1	D1	A2	B2	C2	D2
Number of accesses	51841	51841	51841	51841	80416	79201	81163	76294
Number of identified sequences (sessions)	9342	11052	8256	15019	9342	11052	8256	15019
Number of frequent sequences	61	56	66	46	93	86	98	76

TABLE 6. Incidence of the discovered sequence rules in particular files.

Body	=>	Head	A1	B1	C1	D1	A2	B2	C2	D2
theme2/exploratory data analysis	=>	theme2/exploratory data analysis	1	1	1	1	1	1	1	1
⋮	=>	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
theme2/quiz, theme2/example2	=>	theme2/quiz	1	1	1	0	1	1	1	1
⋮	=>	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
theme3/autotest	=>	course, thema2/example2	0	0	0	0	1	1	1	1
Count of derived sequence rules			38	34	42	22	72	67	76	58
Percent of derived sequence rules (Percent 1's)			42.7	38.2	47.2	24.7	80.9	75.3	85.4	65.2
Percent 0's			57.3	61.8	52.8	75.3	19.1	24.7	14.6	34.8
Cochran Q test			$Q = 158.6498$; $df = 7$; $p < 0.000$							

TABLE 7. Homogeneous groups for incidence of derived rules in examined files.

File	Incidence Mean	1	2	3	4
D1	0.247				****
B1	0.382	****			****
A1	0.427	****			****
C1	0.472	****			
D2	0.652		****		
B2	0.753		****	****	
A2	0.809		****	****	
C2	0.854			****	
Kendall Coefficient of Concordance					0.25465

- almost 5 % of rules in the case of the file (D1) with session identification using STT based on quartiles without path completion (Table 11),
- approximately 10 % of rules in the case of files (A1, B1, C1) using Reference Length method (Table 8, Table 9) and STT based on mean without path completion (Table 10).
- A statistically significant difference was proved in the case of files with session identification using Reference Length method (Table 8, Table 9). The difference consisted of 40-43 new rules found in the files with path completion (A2, B2).
- In the case of files with session identification using STT (Table 10, Table 11) it is also 40-43 new rules, where the statistically significant difference was proved in the number of found rules between the files without and with path completion in favor of files with path completion (C2, D2).

B. QUALITY EVALUATION OF THE EXTRACTED RULES IN THE EXAMINED FILES

The results of sequence patterns analysis can be analyzed more closely considering the portion of each type of found rules. It was required the association rules be not only understandable but also useful. An association rule analysis produced the three common types of rules useful (utilizable, beneficial), trivial and inexplicable [34].

In this case, upon the sequence rules were differentiated the same types of rules. The only requirement (validity assumption) of the use of the chi-square test is sufficiently high expected frequencies [51]. The condition is violated

TABLE 8. Crosstabulations: File A1 x File A2.

A1\A2	0	1	∑
0	8	43	51
	8.99%	48.31%	57.30%
1	9	29	38
	10.11%	32.58%	42.70%
∑	17	72	89
	19.10%	80.90%	100.00%
McNemar (B/C)	*Chi-square* = 20.94231; *df* = 1; *p* = 0.000		

TABLE 9. Crosstabulations: File B1 x File B2.

B1\B2	0	1	∑
0	15	40	55
	16.85%	44.94%	61.80%
1	7	27	34
	7.87%	30.34%	38.20%
∑	22	67	89
	24.72%	75.28%	100.00%
McNemar (B/C)	*Chi-square* = 21.78723; *df* = 1; *p* = 0.000		

TABLE 10. Crosstabulations: File C1 x File C2.

C1\C2	0	1	∑
0	4	43	47
	4.49%	48.31%	52.81%
1	9	33	42
	10.11%	37.08%	47.19%
∑	13	76	89
	14.61%	85.39%	100.00%
McNemar (B/C)	*Chi-square* = 20.94231; *df* = 1; *p* = 0.000		

TABLE 11. Crosstabulations: File D1 x File D2.

D1\D2	0	1	∑
0	27	40	67
	30.34%	44.94%	75.28%
1	4	18	22
	4.49%	20.22%	24.72%
∑	31	58	89
	34.83%	65.17%	100.00%
McNemar (B/C)	*Chi-square* = 27.84091; *df* = 1; *p* = 0.000		

TABLE 12. Crosstabulations: Incidence of rules x Types of rules: File A1.

A1\Type	Useful	Trivial	Inexplicable
0	1	48	2
	4.76 %	88.89 %	14.29 %
1	20	6	12
	95.24 %	11.11 %	85.71 %
∑	21	54	14
	100 %	100 %	100 %
Pearson	*Chi-square* = 56.30237; *df* = 2; *p* = 0.00000		
Con. Coef. C	0.62248		
Cramér's V	0.79537		

if the expected frequencies are lower than 5. The validity assumption of the chi-square test was violated in realized tests. It was a reason why not only the results of Pearson chi-square test were considered, but also the values of calculated contingency coefficient.

Contingency coefficients (Con. Coef. C, Cramér's V) represent the degree of dependency between two nominal variables. The value of Cramér's V coefficient was approximately 0.8 (Table 12). There was a very large dependency between the portion of useful, trivial and inexplicable rules and their occurrence in the set of discovered rules extracted from the file A1. The contingency coefficient was statistically significant. The zero hypothesis (Table 12) was rejected at the 0.1 % significance level, i.e. the portion of useful, trivial and inexplicable rules depended on the identification of sessions by the Reference Length method based on sitemap estimation. In this file, the least inexplicable rules were found (86 %), while 20 useful rules were extracted from the file A1, which represent 95 % of all the found useful rules. The most useful rules were found in the file with the sessions identification based on sitemap estimation (A1).

The value of Cramér's V coefficient (Table 13) was approximately 0.7, where 1 means perfect relationship and 0 no relationship. There was a large dependency among the portion of useful, trivial and inexplicable rules and their occurrence on the set of the discovered rules extracted from the file B1; the contingency coefficient was statistically significant. The zero hypothesis (Table 13) was rejected at the 0.1 % significance level, i.e. the portion of useful, trivial and inexplicable rules depended on the identification of sessions by Reference Length method based on subjective estimation.

The Cramér's V coefficient value (Table 14) was approximately 0.7. There was a large dependency among the portion of useful, trivial and inexplicable rules and their occurrence in the set of discovered rules extracted from the file C1, and the contingency coefficient was statistically significant. In this file, trivial (19 %) and inexplicable (100 %) rules were mostly found.

The Cramér's V coefficient value (Table 15) was approximately 0.77. There was a large dependency among the portion of useful, trivial and inexplicable rules and their occurrence in

TABLE 13. Crosstabulations: Incidence of rules x Types of rules: File B1.

B1\Type	Useful	Trivial	Inexplicable
0	6 28.57 %	48 88.89 %	1 7.14 %
1	15 71.43 %	6 11.11 %	13 92.86 %
∑	21 100 %	54 100 %	14 100 %
Pearson	*Chi-square* = 44.32209; *df* = 2; *p* = 0.00000		
Con. Coef. C	0.57658		
Cramér's V	0.70569		

TABLE 14. Crosstabulations: Incidence of rules x Types of rules: File C1.

C1\Type	Useful	Trivial	Inexplicable
0	3 14.29 %	44 81.48 %	0 0.00 %
1	18 85.71 %	10 18.52 %	14 100.00 %
∑	21 100 %	54 100 %	14 100 %
Pearson	*Chi-square* = 45.98594; *df* = 2; *p* = 0.00000		
Con. Coef. C	0.58367		
Cramér's V	0.71882		

TABLE 15. Crosstabulations: Incidence of rules x Types of rules: File D1.

D1\Type	Useful	Trivial	Inexplicable
0	13 61.90 %	53 98.15 %	1 7.14 %
1	8 38.10 %	1 1.85 %	13 92.86 %
∑	21 100 %	54 100 %	14 100 %
Pearson	*Chi-square* = 52.12257; *df* = 2; *p* = 0.00000		
Con. Coef. C	0.60774		
Cramér's V	0.76528		

the set of discovered rules extracted from the file D1, and the contingency coefficient was statistically significant. In this file, the least useful (38 %) and trivial (2 %) rules were found, while the portion of inexplicable rules has not changed too much (93 %).

In the case of files with the path completion (X2, where X = {A, B, C, D}) the values of contingency coefficients were in range 0.2-0.5, where 1 means perfect relationship and 0 no relationship. There was a small-moderate dependency among the portion of useful, trivial and inexplicable rules and their occurrence in the set of discovered rules extracted from the files with path completion.

The graphs in Fig. 3 – Fig. 6 visualize an interaction frequencies - File X2 x Types of rules. In this case, the curves did not copy mutually; they had different course - which only proved the results of analysis. The most useful rules (57 %) were found in the file A2 and the least in the file D2 (48 %), while the portion of inexplicable rules had not changed too much.

A closer look at the differences between files with and without the path completion (X1 vs. X2) in terms of the dependency on types of the rule (Table 10, Table 11, Table 12, Table 13, see Fig. 3 - Fig. 6) showed increasing number of trivial rules and a decreasing number of useful rules in case of files with path completion (X2) while the portion of inex-

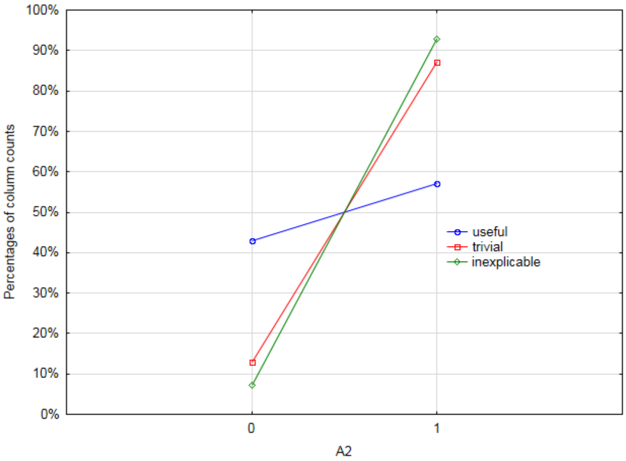


FIGURE 3. Interaction Plot: Incidence of rules x Types of rules: File A2.

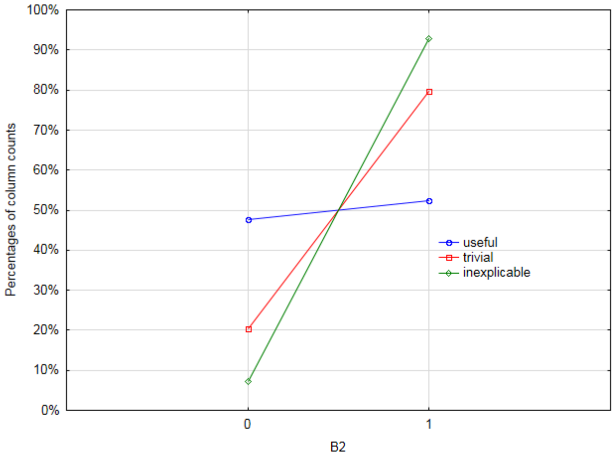


FIGURE 4. Interaction Plot: Incidence of rules x Types of rules: File B2.

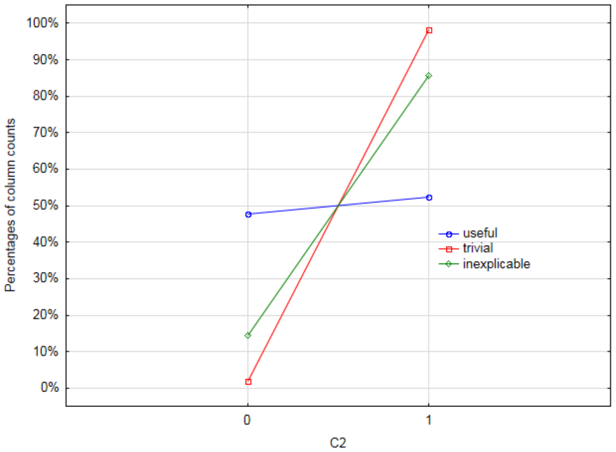


FIGURE 5. Interaction Plot: Incidence of rules x Types of rules: File C2.

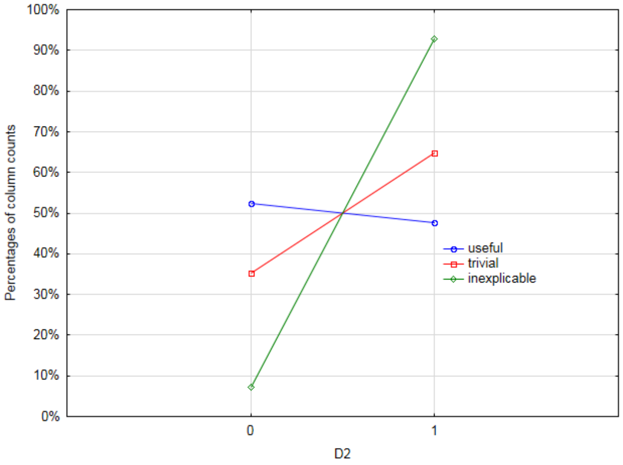


FIGURE 6. Interaction Plot: Incidence of rules x Types of rules: File D2.

plicable rules was similar in both groups of files (X1, X2). The path completion resulted only in increasing of trivial rules.

The quality of found sequence rules was also assessed using two indicators [34]: variables *support* and *confidence*. The results of sequence patterns analysis showed differences not only in the quantity of found rules but also in the quality. Kendall's coefficient of concordance represents the degree of concordance in variable *support* of found rules among examined files. The value of the coefficient (Table 16) was approximately 0.71 while 1 means a perfect concordance and 0 represents discordance.

From the multiple comparisons (Tukey test) only one homogenous group (Table 16) consisting of examined files, were identified regarding the average *support* of found rules. There was not a statistically significant difference in variable *support* of the discovered rules between these files.

Regarding the values of *confidence* of the discovered rules, the differences in the quality of individual files were demonstrated. The coefficient of concordance values (Table 17) was almost 0.21 while 1 means a perfect concordance and 0 represents discordance.

From the multiple comparisons (Tukey test) two homogenous groups (Table 17), consisting of the examined files, were identified regarding the average confidence of the found rules. The first homogenous group consists of files D2, B1, C2, B2, A2, A1, C1 and the second of files D1, D2, B1, C2, B2, A2. There was not a statistically significant difference in the variable *confidence* of discovered rules between these files.

Contrary, statistically significant differences in the level of significance 5 % in the average confidence of found rules were proved between files D1 and A1 as well as D1 and C1. The highest value of *confidence* was achieved in the case of files with session identification using *STT* based on the mean (C1) and using the Reference Length method based on sitemap estimation (A1) without the path completion.

TABLE 16. Homogenous groups for support of derived rules.

File	Support Mean	1
D1	2.631	****
B1	2.785	****
A1	2.819	****
C1	2.821	****
D2	3.417	****
B2	3.521	****
A2	3.557	****
C2	3.562	****
Kendall Coefficient of Concordance		0.71186

TABLE 17. Homogenous groups for the confidence of derived rules.

File	Confidence Mean	1	2
D1	48.724		****
D2	54.124	****	****
B1	54.945	****	****
C2	56.278	****	****
B2	56.354	****	****
A2	56.667	****	****
A1	57.621	****	
C1	58.834	****	
Kendall Coefficient of Concordance		0.20448	

V. DISCUSSION AND CONCLUSIONS

Several data preprocessing techniques and problems related to their application were discussed in this paper. Despite the fact, that the log files came from the VLE, typical techniques for the web data mining and data preprocessing were used. Based on the fact the VLE Moodle records only the activity of registered users in the course, it was not necessary to remove data typically overflowing e-commerce log files. The teachers' accesses and course administrators' accesses were removed because they were not relevant for this research. Similarly, it was not necessary to identify the individual users because the anonymous accesses were not allowed in the examined e-learning course. This fact markedly helped to shorten the data preprocessing phase.

On the other hand, the log file structure, specifically the URL field, represented the main disadvantage. The absence of the domain meant an obstacle in the path completion phase of data preprocessing. It was, therefore, necessary to modify also the results of the sitemap, with the aim to apply a path completion algorithm on these files.

Standard data mining techniques could be used without fail after solving this problem. If the *URL* variable of the log file were in standard HTTP format, it would allow using typical data mining preprocessing techniques without further modifications.

For easier understanding, the list of the assumptions is shown there again, and will be discussed in this section:

1) It was expected that an identification of sessions using the Reference Length method, calculated from a sitemap, would have a significant impact on the quantity of extracted rules.
2) It was expected that an identification of sessions using the Reference Length method, calculated from a sitemap, would have a significant impact on an increasing the portion of useful rules.
3) It was expected that an identification of sessions using the Reference Length method, calculated from a sitemap, would have a significant impact on the quality of extracted rules.
4) It was expected that a path completion would have a significant impact on the quantity of extracted rules.
5) It was expected that a path completion would have a significant impact on an increasing the portion of useful rules.
6) It was expected that a path completion would have a significant impact on the quality of extracted rules.

Comparison of data analysis results elaborated on various levels of data preprocessing regarding quantity and quality of the found rules will be closely described in the next section.

The first assumption concerning the identification of sessions using the Reference Length method, calculated from the sitemap and its impact on the quantity was not proved. The session identification using the Reference Length method based on sitemap estimate has no impact on the quantity of extracted rules in the case of files without and also with path completion. On the contrary, the fourth assumption concerning the path completion and its impact on the quantity was proved. Specifically, it was proved that completing the paths has a significant impact on the quantity of extracted rules. Statistically significant differences in the average incidence of found rules were proved between files without and also with path completion.

In the case of files with path completion, almost 50 % of the new rules were found. On the other hand, the path completion resulted only in an increase of trivial rules. The fifth assumption concerning the path completion and its impact on increasing the portion of useful rules was not proved. On the contrary, the least inexplicable rules and the most useful rules were found in the file with sessions identification based on sitemap estimate. The second assumption concerning the identification of sessions using the Reference Length method, calculated from the sitemap and its impact on increasing the portion of useful rules was proved.

The third assumption concerning the session identification using the Reference Length method based on sitemap estimate and its impact on the quality regarding their basic measures of quality was only proved partially. There was only statistically significant difference in the confidence of discovered rules between examined files. In the case of a file with session identification using the Reference Length method based on sitemap estimate without path completion, the highest value of confidence was achieved. On the other hand, the sixth assumption concerning the path completion and its impact on the quality of extracted rules in term of their basic measures of the quality was not proved.

The experiment also has several didactical outcomes, which can lead to the e-learning course improvement and can help teachers to understand better and to improve their e-learning teaching skills. Since the path completion did not increase useful rules quantity, it can be assumed that the used e-learning course had a well-defined structure and sufficiently intuitive navigation, because the students did not often use the Back button on the browser.

The analysis of discovered sequence rules shown, that some unusual behavior patterns of students can be identified. The rules were classified by the team of human experts into three categories: useful, trivial and inexplicable rules. For instance, activity Task 2 exists in most of the useful rules, which means that the students used it in various activities in the course. One of the possible explanations of the rules with Task 2 is that the students looked for the explanations of terms used in the e-learning course. This finding could inspire the e-learning course developers to create an extra section in the e-learning course, which will summarize all statistical theories or methods used in the course. Similarly, other tips to improve the e-course using the analysis of other rules can be proposed.

The results of the described experiment, together with the previously realized experiments with educational data mentioned in section 3.1 [37]–[39] can be regarded as a basis for the further research in the field of EDM and LA. In this paper, an existing research in the impact of different preprocessing tasks on the quality and quantity of discovered knowledge from the log files stored in contemporary VLEs was extended. The results demonstrate that not all stages of educational data preprocessing are necessary if the data is stored in the VLEs.

Similarly, it was proved that the user session identification using the Reference Length and e-learning course sitemap statistically significantly increased the number of discovered useful sequence rules. The most interesting impact of this result is that the application of the Reference Length method without path completion as well as an e-learning course crawling with the aim to create a sitemap can be automated.

Abovementioned findings could help other educational data mining researchers to choose the suitable steps of educational data preprocessing and easier focus on solving specific educational problems.

In the last few years, researchers have begun to investigate various web log mining methods which allow exploring, visualizing, interpreting and analyzing educational data [27], [28]. The presented outcomes can help developers of VLEs and developers of software tools and plugins for the analysis and presentation of log data stored in VLEs, to automate and accelerate the preprocessing phase of the web log mining. The right decision, which educational data preprocessing tasks are necessary for obtaining relevant results from applied modern EDM and LA methods may lead to the development of real-time reporting tools that monitor students' behavior in the e-learning courses and visualize them in simplified and interactive form.

The future research should, therefore, be focused on improving used methodology, on verifying of the usefulness and effectiveness of other user session and path completion methods on the larger and standardized datasets. Other research should be targeted at an automation of the preprocessing tasks and development of tools, which combine a didactical theory with students' data and serve the needs of all groups of stakeholders. The last research direction can be aimed at the standardization of log formats stored in the VLEs, which may shorten the time required for educational data analysis, including the most time consuming preprocessing stage.

REFERENCES

[1] A. Peña-Ayala, *Educational Data Mining: Applications and Trends*. Heidelberg, Germany: Springer, 2014.

[2] W. Greller and H. Drachsler, "Translating learning into numbers: A generic framework for learning analytics," *Edu. Technol. Soc.*, vol. 15, no. 3, pp. 42–57, Jul. 2012.

[3] G. Siemens, "Learning analytics: Envisioning a research discipline and a domain of practice," in *Proc. 2nd Int. Conf. Learn. Anal. Knowl.*, May 2012, pp. 4–8.

[4] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[5] C. Romero, J. R. Romero, and S. Ventura, "A survey on pre-processing educational data," in *Educational Data Mining* (Studies in Computational Intelligence), vol. 524. Heidelberg, Germany: Springer-Verlag, 2014, pp. 29–64.

[6] G. Siemens and R. S. J. D. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. 2nd Int. Conf. Learn. Anal. Knowl. (LAK)*, vol. 12. 2012, pp. 252–254.

[7] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*. Heidelberg, Germany: Springer, 2014, pp. 61–75.

[8] A. L. Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder, "Design and implementation of a learning analytics toolkit for teachers," *J. Edu. Technol. Soc.*, vol. 15, no. 3, pp. 58–76, 2012.

[9] J. A. Lara, D. Lizcano, M. A. Martínez, J. Pazos, and T. Riera, "A system for knowledge discovery in e-learning environments within the European Higher Education Area—Application to student data from Open University of Madrid, UDIMA," *Comput. Edu.*, vol. 72, pp. 23–36, Mar. 2014.

[10] G. Lavigne, G. G. Ruiz, L. McAnally-Salas, and J. O. Sandoval, "Log analysis in a virtual learning environment for engineering students," *Int. J. Edu. Technol. Higher Edu.*, vol. 12, no. 3, pp. 113–128, 2015.

[11] V. Kovanović, D. Gašević, S. Dawson, S. Joksimović, R. S. Baker, and M. Hatala, "Penetrating the black box of time-on-task estimation," in *Proc. 5th Int. Conf. Learn. Anal. Knowl.*, 2015, pp. 184–193.

[12] V. Chitraa and A. S. Davamani, "A survey on preprocessing methods for web usage data," *Int. J. Comput. Sci. Inf. Secur.*, vol. 7, no. 3, pp. 78–83, 2010.

[13] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in Web-usage analysis," *INFORMS J. Comput.*, vol. 15, no. 2, pp. 171–190, May 2003.

[14] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[15] C. Romero, S. Ventura, R. S. J. D. Baker, and M. Pechenizkiy, *Handbook of Educational Data Mining* (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). London, U.K.: Chapman & Hall, 2010.

[16] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics*. Heidelberg, Germany: Springer, 2014, ch. 4, pp. 61–75.

[17] Z. Papamitsiou and A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Edu. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, 2014.

[18] M. Brown, "Learning analytics: Moving from concept to practice," EDUCAUSE Learning Initiative, Louisville, CO, USA, 2012. [Online]. Available: https://library.educause.edu/resources/2012/7/learning-analytics-moving-from-concept-to-practice

[19] J. A. Larusson and B. White, Eds., *Learning Analytics: From Research to Practice*. Heidelberg, Germany: Springer, 2014.

[20] M. Bienkowski, M. Feng, and B. Means, "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief," SRI Int., Washington, DC, USA, Tech. Rep. ED-04-CO-0040, Task 0010, 2012, pp. 1–57. [Online]. Availbale: https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

[21] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, 2014.

[22] L. Ali, M. Hatala, D. Gašević, and J. Jovanović, "A qualitative evaluation of evolution of a learning analytics tool," *Comput. Edu.*, vol. 58, no. 1, pp. 470–489, 2012.

[23] M. Á. Conde, Á. J. Hérnandez-García, F. García-Peñalvo, M. L. Séin-Echaluce, "Exploring student interactions: Learning analytics tools for student tracking," in *Learning and Collaboration Technologies* (Lecture Notes in Computer Science), vol. 9192, P. Zaphiris and A. Ioannou, Eds. Cham, Switzerland: Springer, 2015, pp. 50–61.

[24] C. G. Marquardt, K. Becker, and D. D. Ruiz, "A pre-processing tool for Web usage mining in the distance education domain," in *Proc. Int. Database Eng. Appl. Symp. (IDEAS)*, 2004, pp. 78–87.

[25] C. Romero, S. Ventura, A. Zafra, and P. de Bra, "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems," *Comput. Edu.*, vol. 53, no. 3, pp. 828–840, 2009.

[26] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Comput. Edu.*, vol. 51, pp. 368–384, Aug. 2008.

[27] I. Dimopoulos, O. Petropoulou, and S. Retalis, "Assessing students' performance using the learning analytics enriched rubrics," in *Proc. ACM Int. Conf. Learn. Anal. Knowl.*, 2013, pp. 195–199.

[28] R. Mazza, M. Bettoni, M. Faré, and L. Mazzola, "MOCLog—Monitoring online courses with log data," in *Proc. 1st Moodle Res. Conf.*, 2012, pp. 132–139.

[29] B. Bakariya, K. K. Mohbey, and G. S. Thakur, "An inclusive survey on data preprocessing methods used in Web usage mining," in *Proc. 7th Int. Conf. Bio-Inspired Comput., Theor. Appl. (BIC-TA)*, vol. 202, J. C. Bansal, P. Singh, K. Deep, M. Pant, and A. Nagar, Eds. Gwalior, India: Springer, 2013, pp. 407–416.

[30] Y. Li, B. Feng, and Q. Mao, "Research on path completion technique in Web usage mining," in *Proc. Int. Symp. Comput. Sci. Comput. Technol. (ISCSCT)*, vol. 1. 2008, pp. 554–559.

[31] C. Ye, J. S. Kinnebrew, and G. Biswas, "Mining and identifying relationships among sequential patterns in multi-feature, hierarchical learning activity data," in *Proc. Edu. Data Mining*, 2014, pp. 389–390.

[32] D. Klocoková, "Integration of heuristics elements in the Web-based environment: Experimental evaluation and usage analysis," *Procedia-Social Behavioral Sci.*, vol. 15, pp. 1010–1014, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042811004095

[33] Z. Abdullah, T. Herawan, H. Chiroma, and M. M. Deris, "A sequential data preprocessing tool for data mining," in *Computational Science and Its Applications* (Lecture Notes in Computer Science), vol. 8581. Cham, Switzerland: Springer, 2014, pp. 734–746.

[34] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Indianapolis, IN, USA: Wiley, 1997.

[35] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1993, pp. 207–216.

[36] *Electronic Statistics Textbook*, StatSoft, Tulsa, OK, USA, 2013.

[37] M. Munk, M. Drlík, J. Kapusta, and D. Munková, "Methodology design for data preparation in the process of discovering patterns of Web users behaviour," *Appl. Math. Inf. Sci.*, vol. 7, no. 1L, pp. 27–36, 2013.

[38] M. Munk and M. Drlík, "Impact of different pre-processing tasks on effective identification of users' behavioral patterns in Web-based educational system," *Procedia Comput. Sci.*, vol. 4, pp. 1640–1649, Jun. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050911002353

[39] M. Munk and M. Drlík, "Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining," in *Digital Information and Communication Technology and Its Applications*, vol. 166, H. Cherifi, J. Zain, and E. El-Qawasmeh, Eds. Heidelberg, Germany: Springer, 2011, pp. 60–74.

[40] T. Huynh and J. Miller, "Empirical observations on the session timeout threshold," *Inf. Process. Manage.*, vol. 45, no. 5, pp. 513–528, 2009.

[41] L. D. Catledge and J. E. Pitkow, "Characterizing browsing strategies in the World-Wide Web," *Comput. Netw. ISDN Syst.*, vol. 27, no. 6, pp. 1065–1073, 1995.

[42] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Berlin, Germany: Springer-Verlag, 2011.

[43] J. Reichel, "Stages of optimization of electronic course computer data analysis," in *Proc. Sci. Iuvenis, Book Sci. Papers*, 2013, pp. 459–464.

[44] J. Reichel and M. Munk, "Reliability/item analysis of statistical literacy tests," in *Proc. 10th Int. Sci. Conf. Distance Learn. Appl. Inform.*, 2014, pp. 483–492.

[45] W. Rice, *Moodle 2.0 E-Learning Course Development*. Birmingham, U.K.: Packt Publishing, 2006.

[46] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," *Knowl. Inf. Syst.*, vol. 1, pp. 5–32, Feb. 1999.

[47] J. Kapusta, M. Munk, and M. Drlík, "Cut-off time calculation for user session identification by reference length," in *Proc. 6th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, 2012, pp. 1–6.

[48] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, Francisco, CA, USA, 1994, pp. 487–499.

[49] J. Han, L. V. S. Lakshmanan, and J. Pei, "Scalable frequent-pattern mining methods: An overview," in *Proc. Tut. Notes 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 5.1–5.61.

[50] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[51] W. L. Hays, *Statistics*. Austin TX, USA: Holt Rinehart and Winston, 1988.

MICHAL MUNK was born in Piestany, Slovakia, in 1979. He received the M.S. degree in mathematics and informatics and the Ph.D. degree in mathematics from Constantine the Philosopher University, Nitra, Slovakia, in 2003 and 2007, respectively. In 2012, he was an Associate Professor in system engineering and informatics with the Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic.

From 2008 to 2012, he was an Assistant Professor with the Computer Science Department, Constantine the Philosopher University, where he is currently an Associate Professor. His research interests include the data analysis, Web mining, and natural language processing.

Dr. Munk has been a member of the Slovak Statistical and Demographic Society since 2005. He was a recipient of the Green Group Award (best paper) of the International Conference on Computational Science 2013 and the Workshop on Computational Finance and Business Intelligence (Barcelona, 2013).

MARTIN DRLíK received the M.S. degree in biophysics from the Faculty of Mathematics, Physics and Computer Science, Comenius University, Bratislava, Slovakia, in 2001, and the Ph.D. degree in theory of computer science education from Constantine the Philosopher University, Nitra, Slovakia, in 2009.

Since 2002, he has been an Assistant Professor with the Computer Science Department, Constantine the Philosopher University. His research interests include the learning analytics, educational data mining, software engineering, and database systems.

Dr. Drlík has been a member of ACM since 2007. He was a recipient of the Green Group Award (best paper) of the International Conference on Computational Science 2013 and the Workshop on Computational Finance and Business Intelligence (Barcelona, 2013).

ĽUBOMíR BENKO received the M.S. degree in applied informatics from the Faculty of Natural Sciences, Constantine the Philosopher University, Nitra, Slovakia, in 2014. He is currently pursuing the Ph.D. degree in applied informatics with the University of Pardubice, Czech Republic.

From 2014 to 2016, he was a Researcher with the Computer Science Department, Constantine the Philosopher University. His research interests are in the field of Web usage mining and big data.

JAROSLAV REICHEL received the M.S. degree in teaching mathematics and informatics from the Faculty of Natural Sciences, Constantine the Philosopher University, Nitra, Slovakia, in 2012, and the Ph.D. degree in theory of teaching mathematics from Constantine the Philosopher University in 2016.

Since 2016, he has been an Assistant Professor with the Computer Science Department, Constantine the Philosopher University. His research interests include the big data, educational data mining, and business intelligence.