IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 7, 2020, accepted January 18, 2021, date of publication February 1, 2021, date of current version February 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056191

Identification of Changes in VLE Stakeholders' Behavior Over Time Using Frequent Patterns Mining

MARTIN DRLIK[1], **MICHAL MUNK**[1,2], **AND JAN SKALKA**[1]

[1]Department of Informatics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, 949 01 Nitra, Slovakia
[2]Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, 532 10 Pardubice, Czech Republic

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

This work was supported in part by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and Slovak Academy of Sciences under Contracts VEGA-1/0776/18, in part by the European Commission ERASMUS+ Programme 2018, under Grant 2018-1-SK01-KA203-046382, and in part by the Scientific Research Project of the Czech Sciences Foundation under Grant 19-15498S.

ABSTRACT Many contemporary studies realized in the Learning Analytics research field provide substantial insights into the virtual learning environment stakeholders' behaviour on single-course or small-scale level. They used different knowledge discovery techniques, including frequent patterns analysis. However, there are only a few studies that have explored the stakeholders' behaviour over a more extended period of several academic years in detail. This article contributes to filling in this gap and provides a novel approach to using homogeneous groups of frequent patterns for identifying the changes in stakeholders' behaviour from the perspective of time. The novelty of this approach lies in fact, that even though the time variable is not directly involved, identification of homogeneous groups of frequent itemsets allows analysis and comparison of the stakeholders' behavioral patterns and their changes over different observed periods. Found homogeneous groups of frequent itemsets, which conform minimal threshold of selected measures, showed, that it is possible to uncover the changes in stakeholders' behaviour throughout the observed longer period. As a result, these homogenous groups of found frequent patterns allow a better understanding of the hidden changes in seasonality or trends in stakeholders' behaviour over several academic years. This article discusses the possible implications of the results and proposed approach in the context of virtual learning environment management and educational content improvement.

INDEX TERMS Association rule analysis, computational and artificial intelligence, learning management systems, predictive models.

I. INTRODUCTION

The availability of different types of web-based educational systems like Learning Management Systems (LMSs) and Massive Open Online Courses (MOOCs) have significantly contributed to that the Computer-Based Education has become an integral and evitable part of the contemporary education at all levels of schools over the last two decades. These systems collect a huge amount of data about all their stakeholders, which are nowadays intensively analyzed using myriads of learning analytics and educational data mining techniques. As a result, many contemporary studies, realized in both mentioned research fields, provide substantial insights into the virtual learning environment (VLE) stakeholders' behaviour on single-course or small-scale level.

Simultaneously, it is quite surprising that despite the availability of data from a longer period, the studies that have explored the changes in stakeholders' behaviour over a larger period of several academic years are still rare. This notion is in line with other researchers in the domain of learning analytics, who stated, that even though the importance of temporality in learning has been long established, it is only recently that serious attention has been paid to explore temporal concepts and data types, analyze methods for exploiting temporal data, techniques for visualizing temporal information, and practical considerations how to effectively use the outcomes of temporal analysis in particular educational contexts. Thus, temporal and sequential nature of the learning

The associate editor coordinating the review of this manuscript and approving it for publication was John Mitchell.

process is receiving increasing interest and suggestions for systematic research, in which the knowledge discovery tasks like time series analysis or data clustering based on different temporal characteristics are mostly applied [1],[2].

The main aim of this article is to contribute to this still new subfield of learning analytics and describe the proposed methodology and the results of the research, which analyzed the data about the VLE stakeholders' activity stored in the form of logs over several academic years using innovative analysis of the identified frequent patterns.

Frequent pattern mining is an important part of data mining, successfully used in many application domains. It is not surprising that it found its application also in education, including learning analytics.

However, the traditional task of frequent itemset mining is to discover groups of items (itemsets) that frequently appear together in transactions made by stakeholder [3]. The comparison and identification of possible similarities or differences in identified frequent itemsets over observed longer period can be challenging because identified frequent itemsets do not carry temporal information and their usefulness or interestingness in the meaning of measures can change.

Therefore, this article provides an alternative approach, how to utilize identified frequent itemsets, which do not carry any temporal information, in the research of behavioral changes of the VLE stakeholders over a longer period. As was emphasized earlier, while the discovering of behavioral patterns in VLE stakeholders' behaviour from the pre-processed data is often based on the analysis of sequences or time series, chosen approach of frequent itemsets analysis examines a set of activities, which the stakeholder has visited or in which he/she has been involved in the e-learning courses in different periods [4].

The proposed approach can be considered unique because it is not focused on the trend estimation or identification of the stakeholders with similar behaviour in the observed period. However, it is focused on the identification of statistically significant changes in the stakeholders' behaviour in the same periods or seasons of several academic years.

Subsequently, the obtained results can be considered useful for different stakeholders of VLE. A teacher can receive the detailed feedback about the learning process, the students' behaviour in the VLE, identify, which types of activities and resources they prefer, what are their typical habits in the meaning of their orientation and navigation between the activities and resources. Simultaneously, the identified homogeneous groups of the frequent itemsets can be successfully used as input to the learning design methods, which deal with the personalization, recommendation or adaptation of the learning content to the target group of stakeholders with the similar behavioral patterns considering their previous activity in the VLE [5].

Moreover, the proposed approach can be further generalized and applied to different application domains because frequent itemsets are not sovereignly related to the education domain. In general, the presented approach can contribute to a better understanding of the principles and mechanisms of the stakeholders' interaction with a more complex information system and their changes over time.

The structure of the paper is as follows. The related work section summarizes the results of the research papers focused on the analysis of frequent itemsets from the VLE logs, description of different approaches to the frequent itemsets analysis as well as their application in related educational domains. The role of the time in this research, an analysis of more extended periods and intervals in given VLE from the seasonality point of view, is also reviewed in detail. The next section introduces the individual steps of the methodology of evaluation of the frequent itemsets in time, which has its roots in more general CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) [6]. Subsequent sections focus on the contribution of the proposed methodology to the analysis of the frequent itemsets, their homogeneous groups, and their possible contribution to the understanding of the seasonality and trends in the VLE stakeholders' behaviour. The presented research is focused on modeling the stakeholders' behaviour by evaluating of frequent itemsets from the longer period in detail. The last section provides a detailed discussion about different facets of the proposed approach, the examples of a possible application of the obtained results and ideas for further research.

II. RELATED WORK

The research presented in this article is based on the close connection of the frequent pattern analysis and temporal analysis, which techniques are separately applied quite often in the learning analytics research field. Their combination, mainly regarding longer period, is rare.

In general, a pattern is a key element of many data knowledge discovery tasks. It represents any type of homogeneity and regularity. Therefore, the pattern is considered a good descriptor of intrinsic and important properties of the data [7]. These patterns should be novel, significant, unexpected, nontrivial and actionable [8].

According to Agrawal *et al.* [9], a frequent itemset can be considered as a set of items, sequences or transactions, which often occur together in the dataset. Frequent itemsets are a form of the frequent pattern [10]. Therefore, frequent itemset also represents some kind of intrinsic characteristic of the investigated dataset. Discovery of all frequent patterns is a common data mining task. Frequent patterns are often further used as components in larger data mining or machine learning tasks [11].

The same authors also introduced the term frequent itemsets mining, which covers the process of frequent itemsets finding and analysis [9]. They considered frequent itemset mining an essential task due to its ability to extract frequently occurring events, patterns or items (symbols or values) in the dataset [12]. Frequent itemsets mining is generally related to descriptive tasks, which try to find comprehensible patterns that represent any interesting behaviour on unlabeled data. The essential descriptive task associated with frequent itemsets mining is association rule mining. Association rule mining and frequent itemset mining have been used interchangeably in different application domains. As a result, they laid the foundations for the current concept of supervised descriptive pattern mining [13], which gathers multiple tasks including contrast set mining, emerging pattern mining, subgroup discovery, class association rules, and exceptional model mining [7].

Whereas many types of data can be represented as transaction databases, this concept has many applications in a wide range of domains [3]. Therefore, it is not surprising that the analysis of frequent itemsets, frequent patterns, as well as association rule analysis, have also been frequently applied in the Learning analytics research field [14].

Many research papers from the Learning Analytics domain focused mainly on discovering frequent itemsets and their evaluation using different measures at the course level. Moreover, they are together with the identified useful association rules traditionally applied for finding possible correlations between individual items of the dataset, for example, in the following applications [15]:

- recommending activities and resources in the VLEs,
- navigating students in a learning path,
- identifying differences between different groups of students,
- discovering interesting relations in data about stakeholders,
- identifying typical students' mistakes,
- optimizing educational content in term of providing relevant content for different groups of stakeholders,
- environment personalization based on an aggregation of similar profiles and domain ontology.

Bazaldua *et al.* [16] also emphasized an essential role of association analysis in Learning Analytics and Educational Data Mining research domains. They considered it the primary method of relationship mining in this research field. The popular techniques of frequent itemsets and association rules mining have been widely used in a variety of contexts [17]. For example, they were used for making recommendations to students, finding common student's mistakes, associations in behavioral patterns of students [18], [19], or finding factors that influence students' performance in e-learning courses [20]. These studies did not consider the time in the meaning of the identification of changes in any trends or seasonality in found frequent itemsets or association rules.

Huang *et al.* [21] applied frequent itemsets analysis for identification of behavioral patterns of stakeholders in online courses. Consequently, they developed several levels of recommendations based on this analysis, which activities and resources should the stakeholders use. These authors similarly did not evaluate frequent itemsets in time. They focused only on finding association rules and assessment of their interestingness.

The VLE stakeholders' behaviour over a short or longer period, and its changes were in the center of interest of several research studies in Learning Analytics and Educational Data Mining domain. However, they used other data mining techniques as are used in this research.

Uzir *et al.* [22] combined agglomerative hierarchical clustering, epistemic network analysis, and process mining to identify and interpret self-regulated learning in terms of the use of learning strategies. They identified four strategy groups derived from three distinct time management tactics and five learning tactics.

Nguyen and Assoc Comp [23] also based his research on temporal information stored in the VLE logfiles. He focused on the role of outliers in an educational context, which can be individual-specific, time-specific, and task-specific.

Saint *et al.* [2] combined simple frequency measures, epistemic network analysis, temporal and stochastic process mining. They concluded that this combination provides with a richer insight into self-regulated learning behaviour of students.

García *et al.* [5] analyzed the results of the case study, in which the association analysis was applied to log files stored in VLE Moodle. The pre-processing phase of the research is the main difference between their approach and the research presented in this paper. The authors aggregated the records at the initial step. Consequently, this step caused a loss of the time dimension. At the same time, this aggregated data about the number of visited activities and achieved grades caused partial disappearance of semantics about the stakeholders' behaviour hidden in the log files. The authors focused on discovering association rules and did not analyze frequent itemsets.

Saleh and Masseglia [24] tried to find itemsets that are frequent over a specific period but would not be extracted by traditional methods since their value of *support* is very low over the whole dataset. They introduced the definition of temporal and solid itemsets, which represent coherent and compact behaviors over specific periods. Simultaneously, they proposed the SIM algorithm for their extraction.

Ale and Rossi [25] proposed their notion of temporal association rules. Their idea consists of extracting itemsets that are frequent over a specific period that is shorter than the whole database. The periods were defined by the lifetime of each item. Therefore, a data mining process for extracting the periods is not necessary since they only depend on the first and last occurrence of each item.

Xie *et al.* [26] proposed a method that takes both the frequency and duration into account. They defined a function for evaluating the importance of events, summarizing them into uniform events according to their semantics. Consequently, they segmented these events using a sliding window to avoid the counting bias issue. As a result, the task of finding temporal characteristics was reduced to mining complex temporally frequent patterns and association rules.

Only several papers published the results of the longitudinal analysis the students' behaviour from the VLE logs. Some of them used a combination of time series techniques [1], neural networks [27] or clustering [28].

Herodotou *et al.* [29] realized longitudinal study to research adoption of predictive learning analytics on different levels.

Mahzoon *et al.* [30] analyzed data over ten years to show how predictions based on temporal and sequential patterns can be utilized within and between terms and academic years.

Boroujeni and Dillenbourg [31] identified different longitudinal students' profiles. Simultaneously, they discovered and tracked latent study patterns using clustering. They proposed pipeline, which allows analysis at different levels of time and activities.

Quan *et al.* [32] investigated in their longitudinal study, the relationship between changes in learning design over time and stakeholders' behaviour. They found that learning designs were able to explain up to 60% of the variability in student online behaviour [33]. However, they tried to predict the students' success or at-risk students. They did not research the changes in students' behaviour, its seasonality or trends over several years.

The similar methodology as is used in this paper was applied in the paper [34]. The authors analyzed web server logs and seasonality in stakeholders' visits of a group of web pages to find behavioral patterns in observed quarters. They tried to estimate the suitable time for publishing information, which the website visitors look for in particular observed periods.

Interestingness quality measures of frequent itemsets and association rules should be used to filter, rank and mainly getting more useful results. These measures can be divided into objective or data-driven (statistical and structural properties of data) and subjective or user-driven (user's preferences and goals) [8]. As a result of the related paper review, interestingness, comprehensibility, and usefulness of the found rule or frequent itemset represent the main qualitative characteristics. However, they are often overlooked due to its subjective nature. For that reason, it is necessary to find other measures, which can be used for expression of the interestingness of the found rule as objectively as possible [35].

Therefore, Han *et al.* proposed to use a more traditional statistical approach, which observes the correlation between the selected attributes [36]. Merceron and Yacef [19] saw the main weakness of the association analysis in the approach, how the useful identified rules were selected. The measure of the usefulness of the rule was evaluated through objective criteria, how the associated items correlated together, not how useful they were from the stakeholder's point of view in the given situation. The authors paid the most significant attention to the strong symmetric associations – association rules, which satisfy several criteria of given metrics. They utilized the found rules for quality improvement and re-design of the e-learning courses. Simultaneously, they recommended involving other metrics of interestingness of found rules, like Jaccard, Laplace, Added Value, Cosine, Phi Coefficient, or Cohen Kappa.

An application of the association analysis in education also has some limitations, which should be considered carefully. The size of the dataset in term of the number of records as well as in terms of unique users can be mentioned as the first limitation. In both cases, the risk of the I. type error increasing consequently. In other words, more false-positive rules or frequent itemsets can be found in the results. Therefore, it is necessary to minimize the number of attributes considered and enrich data about other additional information in this case. Moreover, the results should be verified in the control experiment or research [5].

A limited sample of data, which enters to the research, can also often lead to discovering many trivial or unusual rules even though the values of the parameter minimal support (*minsup*) and minimal confidence (*minconf*) have been chosen appropriately. According to Bazaldua *et al.* [16], involving other additional measures is highly recommended in this case. Eventually, it is possible to add other external constraints, for example, find itemset, which must or must not be included in the preceding or subsequent rule.

Bing *et al.* [37] recommend asking the stakeholders involved in the research to express, which of the found rules they consider useful based on their previous knowledge or experience.

García *et al.* proposed to develop a knowledge base of the rules before the application of association analysis and compare them consequently [5]. A limited count of the itemsets in the previous or consequent rule, application of the frequently used glossary terms, discretization of the numerical values to categorical, which are usually more understandable for a target group of stakeholders, are again considered as the most often used approach for solving these problems [16]. The same authors observed that there is a correlation between the results obtained by the group of experts and suitable chosen combination of metrics mentioned earlier, for example, Jaccard, cosine and support. Therefore, they recommended using this approach for estimating the usefulness of the found rules in cases, when there is a problem to find enough domain experts [38].

On the other hand, if the group of domain experts is already involved into the evaluation of the found rules, the evaluation of the rules and their contribution for the learning analytics domain can be beneficial [39].

III. THE METHODOLOGY OF EVALUATION OF FREQUENT ITEMSETS IN TIME

The methodology for evaluation of frequent itemsets in time is based on more general CRISP-DM methodology. The justified methodology has the following steps [39]:

1. business understanding - the role of frequent patterns analysis in the learning analytics domain, current typical tasks of knowledge discovery in the temporal and predictive learning analytics subfield, the role of time, seasonality and trends in the researched discipline,
2. data gathering from the log files stored in the VLE over several years,
3. data preparation,

   a. mapping attributes from different years to the unified structure,
   b. data cleaning and filtering,
   c. user and session identification,
   d. calculation of derived variables,
   e. feature reduction,
   f. data modification in line with the requirements of the selected algorithm of the association rule analysis,
4. data analysis,
   a. discovering of behavioral patterns of the VLE stakeholders in the meaning of the frequent itemsets during the defined seasons of academic years 2010-2018,
   b. the definition of the suitable threshold values of the selected measures [9],[40],
5. understanding of the results,
6. further pre-processing of identified frequent itemsets, which conform defined minimal values of selected measures,
7. an application of consequent analysis of frequencies of found frequent itemsets, comparison of homogenous groups to compare and evaluate obtained results from the different points of view (academic years, seasons),
8. application of the research outcomes.

IV. RESEARCH BACKGROUND UNDERSTANDING

The e-learning course used in this study dealt with the introductory topics of relational database systems. It was periodically opened during the winter term of eight consequent academic years (2010/11-2017/18) in the VLE of the university. The course was used in the blended form of study.

The course was cloned from the previous version of the course used in the previous academic year. Therefore, the courses contained the same core resources and assignments, and have been gradually enriched by new activities and resources to provide more relevant and up-to-date information, curated educational resources and practical assignments.

The students of the course were motivated to earn grades for activity on the lessons, for solving mandatory and bonus (optional) assignments and quizzes. Moreover, they were asked to create and submit their projects and pass the midterm and final tests. The final grade was calculated as the ratio of the weighted sum of points obtained for individual activities and the total sum of points from all mandatory activities of the course.

Although the structure of the course has slightly changed over time, these changes have not been fundamental. The core activities and resources remain the same over the years. Therefore, they can be considered suitable as the input to the analysis of frequent patterns after application of pre-processing techniques described in the next sections.

As was mention before, modeling of trends and seasonality in VLE Stakeholders' behaviour and analyzing changes in their preferences of different types of activities in different periods using frequent transactions represent the main aim of the article. This kind of problem belongs to the dependency analysis, classification or prediction tasks from the more general knowledge discovery point of view.

V. DATA GATHERING AND PREPARATION

This section deals with the individual steps of the data preparation phase in detail, which could highly influence the final interpretability of the results and their overall contribution to the research and further application.

A. DATA UNDERSTANDING

Data used in the presented research represents the individual accesses of students to the individual parts (modules, resources, interactive and collaborative activities of the e-learning course. Six hundred seventy-two unique students enrolled in the course during this period. The final dataset obtained from the application layer of the VLE, in this case, LMS Moodle, contained 252 340 anonymized records (Table 1).

B. DATA GATHERING

Data, the records about the activity of the VLE stakeholders, is usually stored in the relational database system. Therefore, the preparation of the initial dataset can be created by the application of several SQL scripts directly on the data layer of the VLE. This dataset usually contains the following list of useful attributes: *ID*, *component*, *action*, *target*, *userid*, *courseid*, *objected*, *contextid*, *edulevel*, *crud*, *timecreated*, *ip*. They all come from the table *mdl_logstore_standard_log*, which represents currently preferred internal structure for storing stakeholders' logs in the VLE Moodle. This log system can be easily replaced by the external storage called Learning Record Warehouse, which popularity is rising rapidly today and will provide a dataset in the standardized form for future learning analytics research [41].

The approach for storing logs in the database of VLE has changed during the observed period of eight academic years. As a result, the distribution and overall semantics of the required attributes over the database tables are now less intuitive. They require more complex operations to obtain the dataset in the same structure as was mentioned before.

Therefore, it was more comfortable and more convenient to export the logs directly from the standard application layer of the VLE for this research. Students' personal data was immediately removed. Simultaneously, the user ID was hashed to fulfil the GDPR (General Data Protection Regulation) requirements.

This approach allowed better reproducibility of the research, especially in the cases, when the researchers do not have direct access to the data layer of the VLE.

C. DATA CLEANING AND FILTERING

Considering the main aim of the research to analyze changes in preferences of different types of activities in different periods, only the records about the students were exported using

TABLE 1. The number of students in individual courses/years and the number of accesses to the courses.

	Number of unique users	Labels of columns – academic years								
Labels of Rows – Courses		**10/11**	**11/12**	**12/13**	**13/14**	**14/15**	**15/16**	**16/17**	**17/18**	**Total**
DB10/11	90	22623								22623
DB11/12	92		42623							42623
DB12/13	96			33092	399					33491
DB13/14	89				26088	543	33			26664
DB14/15	98					29079	469			29548
DB15/16	90						32137	432	80	32649
DB16/17	64							37710	234	37944
DB17/18	53								26798	26798
Total	**672**	**22623**	**42623**	**33092**	**26487**	**29622**	**32639**	**38142**	**27112**	**252340**

the application layer of the VLE in the standardized structure of the logs. Moreover, only the records with the value equal to the *participating level* of the attribute *level* were exported, because this level covers all events and actions, which relate to the educational process. The records stored as the reaction to the activity of other stakeholders, for example, during the e-learning course development phase (other) or evaluation phase (teaching level) were not included in the dataset.

As was mentioned earlier, the research analyzed the logs from the period of eight academic years. Almost all software applications and information systems undergo many upgrades during such a long period, which influence the approach, how the logs about the stakeholders' activity are logged. Therefore, the structure of the legacy log files had to be mapped to the current standardized structure of the logs using carefully selected operations.

The data cleaning phase from the unnecessary records represents the next characteristic log mining phase [42]. Unnecessary files, the visits of bots and crawlers are removed in this process. This approach is not required in the case of data cleaning from the VLE because of the required authorization.

On the other hand, all records about the modules, activities and actions, which did not directly relate to the students' activities were removed during this step of data preparation. For example, the records created by the web services, or CLI operations of the administrator were removed in this process. Simultaneously, the records about the experimental modules and plugins as well as modules with a shallow frequency of use also were filtered out in this process.

D. USER SESSION IDENTIFICATION AND PATH RECONSTRUCTION

Each VLE user has a unique identification number. If the role of the guest is enabled in the system, it is necessary to make sure how his/her activity is recorded in the log system of the VLE. Mostly, and this is also the case of the VLE used in this research, guest users can only view the content. They do not interact with others and do not participate in any activities provided in the e-learning course. Therefore, the log file does not contain any record, that would significantly affect the final data file in terms of possible distortion of the user's behaviour. However, the guest role was forbidden in the VLE used in the research. As a result, traditional web mining methods for user identification did not have to be applied.

Considering the requirements of frequent itemsets mining method, which will be used, the next step of the pre-processing phase required identification of individual user sessions in the dataset. The time-oriented heuristic method with Session Timeout Threshold, $STT = 100$ minutes, was selected. The value of STT considers the fact that the VLE was used mostly in the blended form of education and was based on the previous systematic research [39]. The final dataset consisted of 19 154 identified user sessions.

Path reconstruction is the last conventional step of the data preparation phase. Considering the previous research, the structure of the courses, navigation between VLE activities and modules, the path reconstruction is not necessary as was proven in [43]. The reason is that all session would contain the item related to the main page of the course (module *course*), possibly to the student's dashboard. These cases should be removed from the dataset because they do not represent any real activity, but only the necessary navigation step. However, their higher appearance in the dataset can distort the meaning of found frequent itemsets, which would surely overcome the minimum threshold value of defined measures of interestingness.

E. DEFINITION OF DERIVED VARIABLES

The original log file did not provide all variables required by the chosen analytical method. Therefore, the next step of data preparation was focused on the definition of derived variables. The attribute *time* played the most important role in this step. The following derived variables were derived from it:

- *date, time, month, hour*,
- *year, academic-year*, required for the analysis of the users' accesses in accordance with the academic years,
- *week, season,* required for the division of each academic year into several defined periods,
- a *timestamp*, required for the correct identification of the user sessions.

F. FEATURE REDUCTION

The modified dataset contained a relatively large number of attributes with the same granularity in comparison with the original log file. However, several attributes acquired too many different values, which could cause a more problematic interpretation of the found results. Therefore, similarly to other research domains of knowledge discovery, the human domain expert was asked to review:

- whether the attributes, which enter the selected method in the form of independent variables, have enough semantics for correct interpretability of the results and were comprehensible for the target group of users,
- whether the values of attributes should be replaced by the abstractions/aggregations with better semantics.

Subsequently, several other pre-processing steps were identified as required as a result of this review. This step of the data preparation phase is often the most time-consuming operation, which can be only partially automated. The reason is that the dataset must be repeatedly browsed, filtered, sorted and re-calculated with the help of the expert. The research described in this paper required the following modifications:

The values of the attribute week (1-52) were aggregated to the several distinct seasons of the academic year (season):

- 1 - the season before and at the beginning of the academic year (week 38-44),
- 2 - the season till the end of the first thematic part of the course, which finishes with the midterm test or evaluation activity (week 45-49),
- 3 - season till the end of the term, which correspond with the second thematic part of the course (week 50-2),
- 4 - the season for final exams (week 3-8),
- 0 - the season created for the completeness, which covers all students' records created after the official end of the course. The students have access to their courses until the end of the academic year (week 9- 37).

The original attribute context, which related to the individual types of VLE modules, was combined with the attribute detail and finally replaced by the variable type. As a result, this step unified potential changes in the titles of the same items in different cycles of the e-learning course. Simultaneously, the original information about the purpose of the module or resource used in the e-learning course was replaced by the information, what is the purpose and aim given module or activity in the course. Considering the fact, that several new resources and activities were added to the courses during the observed academic years, this approach eliminated the influence of the absolute frequencies of the visits of the particular resources and activities. At the same time, the real aim or contribution of the module was highlighted.

Three artificial (dummy) variables were defined as follows

- *required* (0,1) - determines, if the type of activity was mandatory,
- *interactive* (0,1) - defines, if this type of activity requires interactivity with the student,
- *priority* (1,2,3) - determines, to what extend the domain expert assumes, this activity is vital for the successful passing of the course by the student.

Finally, the aggregated variable *type-detail* in the form of type-detail-required-interactive-priority was created. This combination of the attributes reduced the original number of the unique items (235) available for the students enrolled in the e-learning courses. As a result, the new variable can acquire 61 different items. (Figure 1).

context	***type***	***type-detail***
Assignment	*activity at lesson*	*other-0-0-3*
Book	*additional resource*	*feedback-0-0-2*
Course	*bonus activity*	*project-1-0-1*
Database	*Feedback*	*bonus activity-1-1-3*
Discussion	*final exam*	*midterm test-1-0-1*
Feedback	*individual assignment*	*software-0-0-3*
Folder	*Instruction*	*project-1-1-3*
Glossary	*Lecture*	*project-0-0-3*
Label	*midterm test*	*software-1-1-3*
Others	*Other*	*additional resource-1-0-1*
Page	*Project*	*lecture-1-0-3*
Quiz	*Software*	*lecture-1-0-1*
Resource	*solved example*	*solved example-1-1-1*
Survey		*final exam-1-1-1*
URL		*midterm test-1-1-1*
Workshop		*...*

FIGURE 1. Feature reduction and mapping.

The final pre-processed dataset contained the following three new variables:

- *academic-year*,
- *season*,
- *type-detail*.

VI. MODEL DESCRIPTION

Finding frequent itemsets can be seen as a simplification of the unsupervised learning problem called "mode finding" or "bump hunting". Each item is seen as a variable in this case. The goal is to find prototype values so that the probability density evaluated at these values is sufficiently large. However, whereas a probability estimation is unreliable and computationally too expensive in real situations, frequent itemsets are used instead of probability estimation [44].

The problem of frequent itemset mining is formally defined as follows [45].

Let there be a set of items (symbols) $I = \{i_1, i_2, \ldots i_m\}$. A transaction database $D = \{T_1, T_2, \ldots T_n\}$ is a set of transactions such that each transaction $T_q \subseteq I (1 \leq q \leq m)$ is a set of distinct items and each transaction T_q has a unique identifier q called transaction identifier.

An itemset X is a set of items such that $X \subseteq I$. Let the notation $|X|$ denote the set the number of items in an itemset X (cardinality). An itemset X is said to be of length k or a k-itemset if it contains k items ($|X| = k$).

The goal of itemset mining is to discover interesting itemsets in a transaction database. In general, various measures can be used to assess the interestingness of patterns in frequent itemset mining.

The interestingness of a given itemset is traditionally defined by a measure called the *support*. The value of measure *support* for an itemset is given by a proportion of records in the transactions data set that have the itemset. More precisely, measure *absolute support* of an itemset X in a database D is denoted as $sup(X)$ and it is defined as the number of transactions containing X, thus $\sup(X) = \left|\{T | X \subseteq T \bigwedge T \in D\right|\}$. Other authors prefer to define measure *support* of an itemset X as a ratio (*relative support*), which is denoted as $relSup\,(X) = \sup\,(X) \,/\, |D|$.

As was mentioned earlier, the task of frequent itemset mining is a process of discovering all frequent itemsets in a given transaction database. An itemset X is considered frequent if it has a *support* that is greater or equal than a given minimum support threshold *minsup*, set explicitly (i.e. $sup(X) \geq minsup$) [3].

Another interesting measure, used frequently in case of association rules analysis, is *confidence*. It can be calculated as follows

$$conf\ (X \Rightarrow Y) = \frac{sup\left(X \bigcup Y\right)}{sup(X)},$$

where an implication of $X \Rightarrow Y$ is defined as rule, $X, Y \subseteq IandX \cap Y = \emptyset$. Although this measure will not be used directly in the paper, it can be used for calculating of measure *lift*. It is defined as follows

$$lift\ (X \Rightarrow Y) = \frac{sup\left(X \bigcup Y\right)}{sup\ (X)\ sup\ (Y)}$$

or using confidence

$$lift\ (X \Rightarrow Y) = \frac{conf\ (X \Rightarrow Y)}{sup(Y)}.$$

The measure lift can be interpreted as the deviation of the support of the whole rule from the support expected under independence, given the supports of the X and Y. Greater lift values indicate stronger associations [46].

It is necessary to design algorithms that avoid exploring the search space of all possible itemsets and process each itemset in the search space as efficiently as possible. Several efficient algorithms have been proposed, like Apriori, FP-Growth, or Eclat. The Apriori algorithm, a horizontal breadth-first search algorithm [3], together with a structured tree procedure that requires only one pass through data, has been finally selected. This algorithm is enclosed in the association rule analysis package STATISTICA Sequence, Association, & Link Analysis [34].

VII. ANALYSIS OF SEASONALITY BASED ON FREQUENT ITEMSETS

This section provides the example, how the frequent itemsets found in individual seasons of the academic years can be interpreted and useful. Its main aim is to show that:

- adding derived variables can make the interpretation of the found frequent itemsets clearer with preserving the semantics,
- identified frequent 1-itemsets provide the option, how to effectively compare the course designer's or teacher's intended purpose of the educational resources and interactive activities with their real use in different seasons of the academic years,
- identified frequent 2-itemsets can uncover unseen relations between pairs of activities or resources, which the students prefer to use together.

Even though this qualitative analysis of the results based on the selected frequent itemset mining technique and frequent itemsets visualization can provide an interesting insight into the individual periods, their comparison and identification of possible changes in stakeholder's behaviour or preferences in selecting particular activities remain challenging. Therefore, a proposal, how to evaluate found frequent itemsets over different periods, will be introduced in the next section.

A. SEASON 1–BEGINNING OF THE TERM

Season 1, which covers the beginning of the academic years 2010/11 – 2017/18, is used as an example, how to understand the results of the analysis.

The activity *bonus activity-1-1-3* with *support* 58% represents the most visited type of activity in academic year 10/11. The following two types of activities *lecture-1-0-1* and *software-0-0-*3 occurred with the probability greater than 30%. On the other hand, the last two activity types, *individual assignment-1-1-3* and *additional resources-1-0-1*, which fulfilled the condition $minsup = 5\%$, were included only in each 15th session from the observed period. The most visited pair of activities, *bonus activity-1-1-3* and *lecture-1-0-1* reached the *support* greater than 20%, and *project-1-0-1* and *bonus activity-1-1-3* (18%) respectively.

A positive correlation ($lift > 1$) between the pairs of activity types was found in several cases. In other words, the students visited these types of activities more often together than individually in all identified sessions. The largest correlation was found for the following two pairs, *solved example-1-1-1 => solved example-1-0-3* and *lecture-1-0-3 => additional resource-0-0-2* ($lift = 3.2$). The probability that the student visited the activity type *solved example-1-1-1* and the activity *solved example-1-0-3* during one session is, in this case, 3.2-times greater than the probability, that the activity type *solved example-1-0-3* occurs in the randomly selected session.

Quantitative evaluation of the results from the first observed season can be summarized as follows (Figure 2):

- The list of the most visited 1-itemsets shows that the mandatory use of the activity relates closely to the frequency of its use during an individual user session.
- It is surprising at first glance that half of the resources and activities with the lowest priority is involved in each fifth observed session.
- Fourteen itemsets, which relate to different types of activities, have values of *support* greater than 5%. Two-thirds of them were mandatory, and two-thirds did not require direct interactivity of the user. Moreover, almost one-third of them had the lowest priority defined by the course designer or teacher.
- At the same time, frequent 2-itemsets uncover other interesting findings – *bonus activity* is involved in all the most visited 2-itemsets. It indicates that the student often uses previously solved examples or additional resource as the primary source of inspiration/help, how to solve the given assignment.

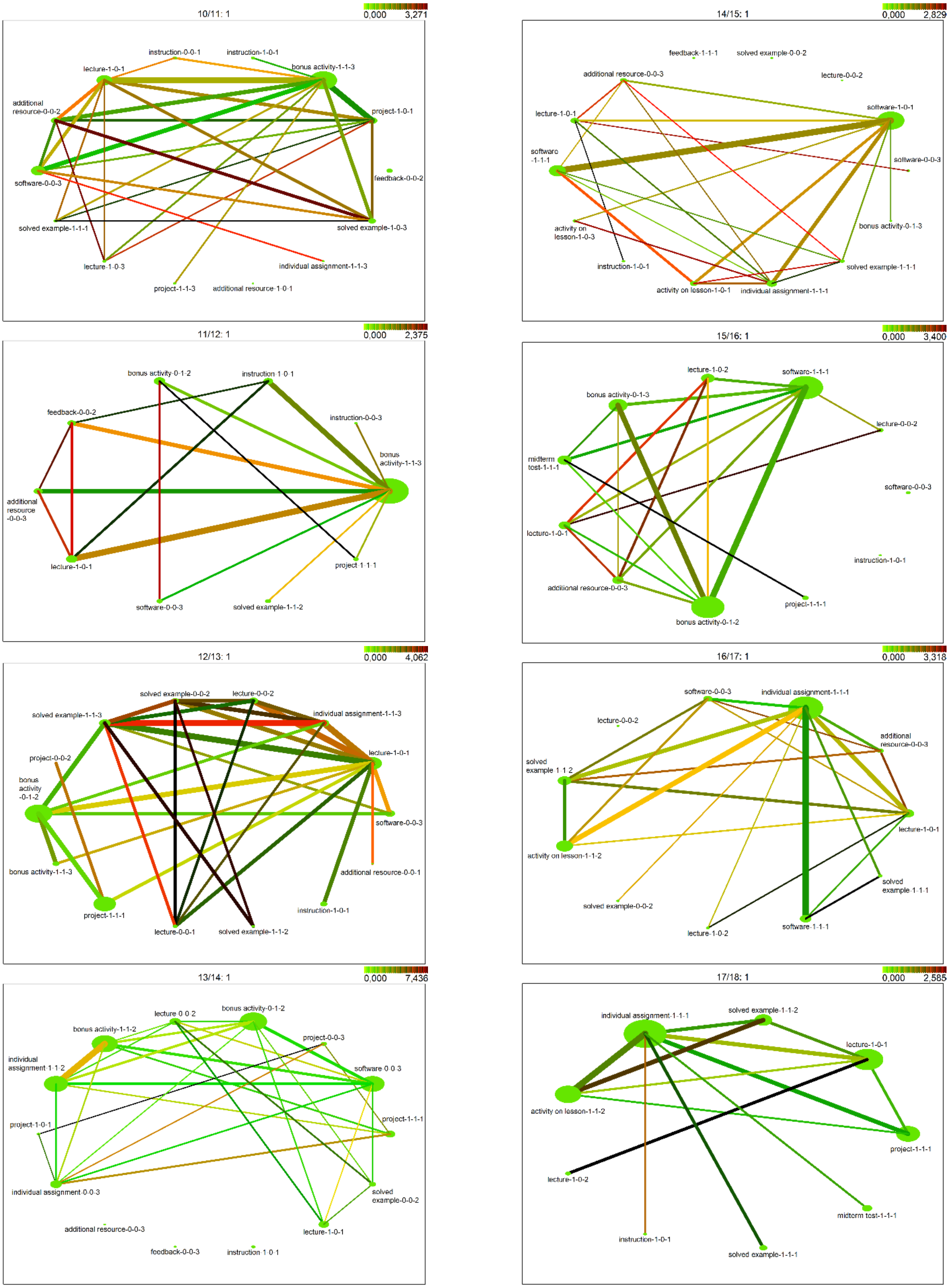

FIGURE 2. The visualization of the first observed season during the academic years 2010-2017 using web chart.

- The characteristic *lift* shows that the activity types *solved example* and *additional resource* have always threefold higher probability that they will create together a part of the session than if they were random. This finding is in line with the requirements of their use as an additional source of information.

An important position of the bonus activities in the first examined period, as well as a positive correlation between additional resources, depicts Figure 2.

Type activity *bonus activity-1-1-3* reached the *support* 76% in academic year 11/12. The *support* of the other two types of activities (*bonus activity-0-1-2*, *lecture-1-0-1*) was much smaller. These types of activities were present in each fifth session. The students used during the session a pair of activity types with the probability of 18%.

Further analysis of the characteristics *lift* and the comparison with the same season of the previous academic year lead to the finding that the positive correlation between the activities oscillated between the values 2.2 – 2.3.

A quantitative evaluation indicates that the stakeholders focused on solving bonus assignments in this academic year. However, after a further detailed analysis of the related logs was clear that the activity at the lesson was mistakenly included in this activity type. These findings also influenced the obtained number of frequent 2-itemsets, in which all identified pairs of activity types with the *support* greater than 10% also contained the bonus activities.

The fact, that eight types of activities reached the *support* greater than 16% can be considered the main characteristic of the academic year 12/13 (*bonus activity-0-1-2* 41%, *project-1-1-1* 34%, *lecture-1-0-1* 29%, *solved example-1-1-3* 19%, *individual assignment-1-1-3* 18%, *software-0-0-3* 17%, *solved example-0-0-2* 16%).

The same types of activities were included in the identified frequent 2-itemsets (*individual assignment-1-1-3*, *solved example-1-1-3*), (*lecture-1-0-1*, *lecture-0-0-2*), (*lecture-1-0-1*, *individual assignment-1-1-3*), (*lecture-1-0-1*, *solved example-1-1-3*). Moreover, these pairs occurred in each tenth session on average in the observed period.

Common, four times higher probability of the occurrence of the pair *solved examples* and *study materials* could indicate a preferred learning approach of the students. This finding is in line with the fact that the data came from the courses, which emphasize practical skills and solving assignments.

A global view on frequent 1-itemsets, as well as on pairs with *lift* > 2 uncovered the finding that a new type of activity *individual assignment* (*support* = 18%) appeared in the structure of the course. This activity type allowed to specify the differences between mandatory and *bonus assignments* (*support* = 41%). Simultaneously, a web chart (Figure 2) depicts a tendency in students' preferences for using several different types of activities.

Four types of activities reached the *support* greater than 34% in the next year 13/14 (*bonus activity-0-1-2*, *bonus activity-1-1-2*, *software-0-0-3*, *individual assignment-1-1-2*). Six different pairs of activities had *support* value greater than 10%. Finally, four times greater probability of common occurrence (*lift* > 4) reached six frequent 2-itemsets.

Mandatory and bonus practical types of activities belonged to the most visited. Unlike the previous observed period, frequent 1-itemsets of type *individual assignment* were included in the comparable number of sessions as *bonus activities* (*support* = 34%).

Surprisingly, activities included in the project were also identified in the frequent itemsets. Considering the characteristic *lift* > 3, it is possible to assume that the project tasks require a common use of several types of activities.

The students visited mostly the activity types *software-1-0-1* (58%), *software-1-1-1* (37%) in the first season of year 14/15. Approximately in each fifth session contained 1-itemsets *individual assignment-1-1-1* (23%), *lecture-1-0-1* (19%), as well as *activity on lesson-1-0-1* (18%). These items also created the most frequently visited pairs of activity types (*software-1-0-1*, *software-1-1-1*) with *support* = 28%, (*software-1-0-1*, *individual assignment-1-1-1*) 17%, (*software-1-0-1*, *activity on lesson-1-0-1*) (13%), and finally (*software-1-1-1*, *activity on lesson-1-0-1*) (12%). It is interesting to note that although the positive correlation between the items decreased, the number of activity types, which occurred together in session, increased up to 14 (*lift* > 2).

The structure of the frequent 1-itemsets remained similar to the previous year. Activity type *software* related closely to a new, quite often visited *activity on lesson*. It can be assumed that the *activity on lesson* was conditional by using software and the related guides or tutorials. This statement is in line with the probability of the common occurrence of the pairs of activities, where the activity software always appeared in the sessions. In general, the overall decrease of both characteristics (*support*, *lift*), as well as frequencies of the items in sessions, was worthy of notice.

The preferences in selecting a particular activity type stayed remained. Seven items reached the *support* greater than 15%: (*software-1-1-1*) 44%, (*bonus activity-0-1-2*) 42%, (*bonus activity-0-1-3*) 25%, (*midterm test-1-1-1*) 18%, (*lecture-1-0-2*) 18%, (*lecture-1-0-1*) 16%, (*additional resource-0-0-3*) 16%. These activities created also the most visited pairs (*software-1-1-1*, *bonus activity-0-1-2*) with *support* = 21%, and *support* = 18% (*bonus activity-0-1-3*, *bonus activity-0-1-2*). Only two pairs occurred in the sessions more frequently together than in the case of their random occurrence (*lift* > 3,4) (*midterm test-1-1-1* ==> *project-1-1-1*), (*lecture-0-0-2* ==> *lecture-1-0-1*).

An activity (*individual assignment-1-1-1*) was present in 70% o and (*activity on lesson-1-1-2*) of all sessions in academic year 16/17. Other three types of activities were included in almost one-quarter of all sessions (*solved example-1-1-2*) 27%, (*lecture-1-0-1*) 24% and (*software-1-1-1*) 23%.

The most frequently visited 2-itemset with *support* around 20% were (*individual assignment-1-1-1*, *software-1-1-1*), (*individual assignment-1-1-1*, *activity on lesson-1-1-2*). However, only two pairs of activity types occurred more times together (*lift* > 3) in sessions than in case their random occurrence.

The structure of the visited activities returned to the normal. Surprisingly, about 70% of the identified transactions

TABLE 2. **The most frequent 1-itemsets found in the second observed season during the academic years 2010-2017.**

frequent 1-itemset	year	*support* (%)
(bonus activity-1-1-3)	**10/11**	**37.11**
(software-1-1-3)	10/11	28.67
(lecture-1-0-3)	10/11	22.89
(project-1-1-1)	**11/12**	**43.11**
(bonus activity-1-1-3)	11/12	35.82
(midterm test-1-1-1)	11/12	20.42
(instruction-1-0-1)	11/12	18.96
(project-1-1-1)	**12/13**	**71.65**
(software-1-0-1)	12/13	23.37
(project-0-0-2)	12/13	21.13
(project-1-1-1)	**13/14**	**62.20**
(software-0-0-3)	13/14	29.71
(project-1-0-1)	13/14	29.29
(midterm test-1-1-1)	**14/15**	**41.76**
(individual assignment-1-1-1)	14/15	24.13
(software-0-0-3)	14/15	22.96
(project-1-1-1)	14/15	20.47
(bonus activity-0-1-3)	14/15	19.80
(project-1-1-1)	**15/16**	**43.24**
(midterm test-1-1-1)	15/16	39.75
(software-0-0-3)	15/16	24.96
(individual assignment-1-1-1)	15/16	20.83
(bonus activity-0-1-3)	**16/17**	**51.93**
(project-1-1-1)	16/17	35.26
(individual assignment-1-1-1)	16/17	34.71
(individual assignment-1-1-1)	**17/18**	**55.67**
(project-1-1-1)	17/18	44.22
(software-1-1-1)	17/18	31.65
(midterm test-1-1-1)	17/18	24.13

contained the activity type *individual assignment* as well as other practically oriented activities.

The ongoing trend in selecting some kinds of activities is also visible in the last observed year 17/18: (*individual assignment-1-1-1*) 44%, (*lecture-1-0-1*) 34%, (*activity on lesson-1-1-2*) 29%, (*project-1-1-1*) 26% and (*solved example-1-1-2*) 20%.

The most visited pairs of activities (*individual assignment-1-1-1*, *activity on lesson-1-1-2*) reached *support* 16%. Simultaneously, the decrease in the highest values of the characteristic *lift* is visible. In other words, only two pairs had the probability of the common access greater than 2.

The absence of the greater count of common appearances of activities during the individual session is probably caused by the continual decreasing interest in course activities. The usage of the course is gradually limited to the accesses to the mandatory activities, like assignments or quizzes. The learning process often happened outside the VLE.

The similar qualitative analysis of the found frequent 1-itemsets and 2-itemsets can be realized for all remaining seasons.

B. SEASON 2–THE END OF THE FIRST THEMATIC TOPIC AND MIDTERM TEST

The most frequently visited itemsets identified in the seconds season with *support* $>$ 20% are summarized in Table 2.

TABLE 3. **The most frequent 2-itemsets found in the second observed season throughout the academic years 2010-2017.**

frequent 2-itemsets	year	*support* (%)
(midterm test-1-0-1, bonus activity-1-1-3)	10/11	16.87
(bonus activity-1-1-3, project-0-0-3)	10/11	16.39
(midterm test-1-0-1, project-0-0-3)	10/11	16.14
(project-1-1-1, bonus activity-1-1-3)	11/12	13.13
(project-1-1-1, project-0-0-2)	11/12	10.53
(project-1-1-1, project-0-0-2)	12/13	18.38
(project-1-0-2, project-1-1-1)	12/13	10.14
(project-1-1-1, project-1-0-1)	13/14	24.69
(project-1-1-1, project-0-0-3)	13/14	15.48
(project-1-0-1, project-0-0-3)	13/14	15.20
(software-1-1-1, midterm test-1-1-1)	14/15	13.81
(midterm test-1-1-1, project-1-1-1)	15/16	13.20
(software-0-0-3, project-1-1-1)	15/16	13.20
(project-1-1-1, bonus activity-0-1-3)	16/17	18.18
(individual assignment-1-1-1, software-1-1-1)	17/18	26.93
(project-1-1-1, individual assignment-1-1-1)	17/18	14.48

The table shows that the type of activity with the highest occurrence has changed throughout the years.

Activity type *project* dominated in this season from the qualitative point of view. It decreased again in 14/15. The students limited their activity in courses mainly on mandatory activities (based on the expert domain decision) and activities, which closely related to the project. These activities were quite popular, even though the domain expert assigned them a lower priority.

Although the course provided an extensive portfolio of other types of activities and resources, its real use was low. The year 12/13 can be considered in this context as extreme (Table 2) because up to 70% of all sessions from this season contained the activity *project*. In contrast, the *support* of other activities was small in comparison with other years.

A short view of the identified frequent 2-itemsets (Table 3) uncovers interesting stakeholders' behaviour. They often used additional study materials and other activities with low priority defined by the expert during the solving assigned mandatory activities. As can be seen in Table 3 as an example, the stakeholders used except the activity type *project* or *midterm test* other *bonus activities* and *individual assignments*. These types of activities created the most frequent itemsets identified in this observed season.

Table 4 shows that the value of *lift* continually decreased. It means that the probability of the visit of several types of activities during one session decreased. In other words, the e-learning course played the role of the storage of tasks and assignments. It lost the position of the center of knowledge and curated content. The students visited the course with the narrow intent. They left the course immediately after finishing the assigned task.

C. SEASON 3–END OF THE SECOND TOPIC

The second learning topic of the course (Table 5) can be again characterized by the activity types project and other project

TABLE 4. Association rules with the highest lift value found in the second observed season throughout the several academic years.

body => head	year	*lift*
project-0-0-3 ==> midterm test-1-0-1	10/11	5.8414
individual assignment-1-1-3 ==> solved example-1-0-3	10/11	4.7572
project-0-0-2 ==> project-1-0-2	11/12	7.6658
project-0-0-2 ==> project-1-0-2	12/13	4.2738
solved example-0-0-2 ==> lecture-0-0-2	13/14	4.7815
individual assignment-0-1-2 ==> project-1-1-1	14/15	4.6641
software-1-1-1 ==> individual assignment-1-1-1	15/16	3.2010
software-1-1-1 ==> individual assignment-1-1-1	16/17	1.8006
software-1-1-1 ==> lecture-1-0-2	17/18	2.2979

TABLE 5. The most frequent 1-itemsets found in the third observed season during the academic years 2010-2017.

frequent 1-itemset	year	*support* (%)
(bonus activity-1-1-3)	**10/11**	**47.85**
(lecture-1-0-1)	10/11	24.54
(lecture-1-0-3)	10/11	23.93
(project-1-1-1)	**11/12**	**62.54**
(bonus activity-1-1-3)	11/12	38.26
(project-0-0-2)	11/12	21.54
(bonus activity-0-1-2)	**12/13**	**47.48**
(project-1-1-1)	**12/13**	**46.22**
(lecture-1-0-1)	12/13	19.33
(project-1-1-1)	**13/14**	**70.93**
(lecture-1-0-1)	13/14	23.26
(project-1-0-1)	13/14	20.93
(project-1-1-1)	**14/15**	**44.68**
(project-0-0-2)	14/15	23.88
(lecture-1-0-1)	14/15	19.39
(project-1-1-1)	**15/16**	**72.88**
(lecture-1-0-1)	15/16	12.50
(project-1-1-1)	**16/17**	**76.62**
(final exam-1-1-1)	16/17	18.18
(individual assignment-1-1-1)	16/17	17.80
(project-1-1-1)	**17/18**	**72.25**
(midterm test-1-1-1)	17/18	40.00

related types of activities. While the bonus activity occurred almost in half of all session in the first academic years, the frequency of visits to other types of activities dramatically decreased, and the *support* is lower than 10%. More than 70% of all transactions realized in the last three years contained the activity type *project*, whereas other types of activities were much less visited.

Table 6, which visualizes the most frequent pairs of activities, confirms the observations mentioned above. Mandatory activities related to the project prevailed in all observed years. The common use of *bonus activities* and activity type *project* can be explained by the assumption that these activities provide a suitable form of teacher's feedback, show analogy of the solved problem and guide student to find a correct solution of the project tasks.

Table 7 depicts associative rules with the highest value of *lift*. The probability of visit to these frequent itemsets during a session is greater than the probability of individual one. Moreover, the decrease of the *lift* values throughout the years is notable with the maximal decrease in academic year 15/16. The observed sessions from this period did not contain any pair of frequent itemsets with *lift* > 1. The academic year 12/13 represents the second extreme because the pair midterm *test-1-1-1 ==> final exam-1-1-1* appeared twelve times more often together in session (*lift* $= 12$) than would be their random occurrence.

TABLE 6. The most frequent 2-itemsets in the third season in the observed period of academic years 2010-2017.

frequent 2-itemsets	year	*support* (%)
(additional resource-1-0-1, lecture-1-0-3)	10/11	12.88
(project-1-1-1, bonus activity-1-1-3)	11/12	17.68
(project-1-1-1, project-0-0-2)	11/12	17.20
(project-1-1-1, bonus activity-0-1-2)	12/13	15.97
(bonus activity-0-1-2, bonus activity-1-1-2)	12/13	13.24
(project-1-1-1, project-1-0-1)	13/14	20.35
(project-1-1-1, project-0-0-3)	13/14	15.70
(project-1-1-1, project-0-0-2)	14/15	21.04
(project-1-1-1, project-0-0-3)	14/15	13.24
(project-1-1-1, midterm test-1-1-1)	15/16	6.60
(individual assignment-1-1-1, project-1-1-1)	16/17	10.95
(project-1-1-1, midterm test-1-1-1)	17/18	25.75

TABLE 7. Associative rules with the highest value of the lift found in the third observed season.

body => head	year	lift
solved example-1-0-3 ==> additional resource-0-0-2	10/11	6.5859
solved example-1-1-2 ==> solved example-1-1-3	11/12	8.9684
solved example-1-1-3 ==> individual assignment-1-1-3	11/12	6.0645
midterm test-1-1-1 ==> final exam-1-1-1	12/13	12.094
project-1-0-2 ==> project-0-0-2	12/13	7.3579
software-1-1-1 ==> bonus activity-1-1-2	12/13	6.4464
solved example-1-1-3 ==> individual assignment-1-1-3	12/13	5.8947
lecture-0-0-2 ==> solved example-0-0-2	13/14	6.3067
feedback-0-0-3 ==> project-0-0-3	13/14	4.1061
software-1-0-1 ==> individual assignment-1-1-1	14/15	4.9236
lecture-1-0-2 ==> lecture-1-0-1	14/15	4.1665
-	**15/16**	
bonus activity-0-1-3 ==> individual assignment-1-1-1	16/17	3.5091
individual assignment-1-1-1 ==> final exam-1-1-1	17/18	3.9308
lecture-1-0-1 ==> individual assignment-1-1-1	17/18	2.2642

D. SEASON 4–PERIOD OF FINAL EXAMS

The nature of tasks, which should be solved by the students, differs in the following two seasons. It could be expected that the *final-exam-1-1-1* is the most frequently visited type of activity with the *support* in the interval 40-60%.

All itemsets with the *support* $> 15\%$ were mandatory and required some kind of interactivity, except for the activity type *lecture-1-0-1*. Surprisingly, this activity did not belong to the most visited activities in the period of final exams. At the same time, the value of *support* for other frequent itemsets was very low.

Figure 3 visualizes the relation between frequent 1-itemsets and 2-itemsets. While the frequent 1-itemsets are present in half of all sessions, frequent 2-itemsets

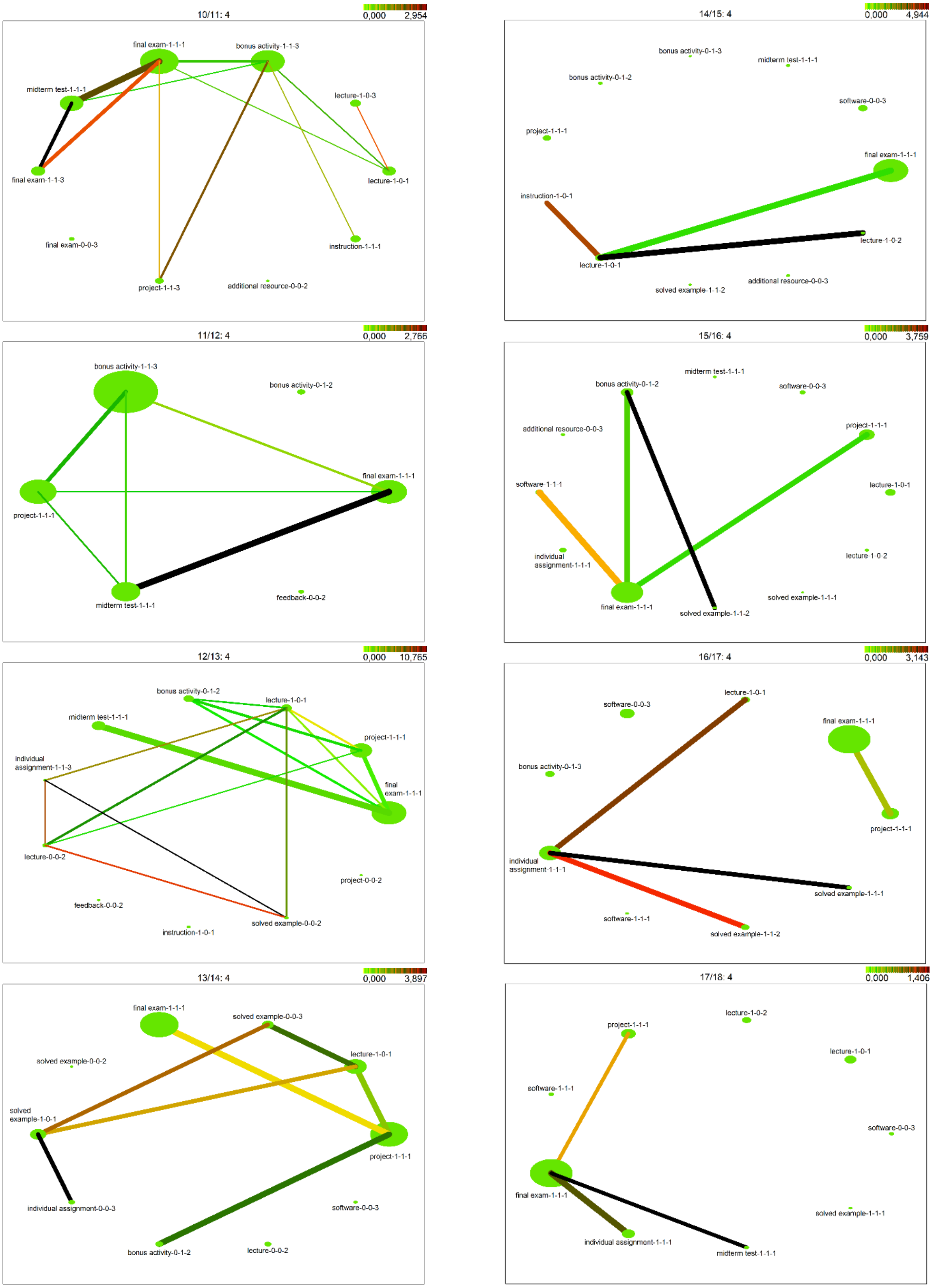

FIGURE 3. Web graphs of the final exam period in the academic years 2010-2017.

have become rarer from the academic year 13/14. In other words, a decrease in the number of occurrences of frequent 2-itemsets is notable mainly in the last four academic years.

The highest positive correlation between two items was reached in 12/13 (*solved example-0-0-2* ==> *individual assignment-1-1-3*, *lift* = 10.7, *solved example-0-0-2* ==>

frequent itemsets	10/11:1	10/11:2	10/11:3	10/11:4	…	17/18:2	17/18:3	17/18:4
(activity on lesson-0-1-1)	*0*	*0*	*1*	*0*		*1*	*0*	*0*
(activity on lesson-0-1-1, bonus activity-1-1-3)	*0*	*0*	*1*	*0*		*0*	*0*	*0*
(activity on lesson-1-0-1)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(activity on lesson-1-0-1, individual assignment-1-1-1)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(activity on lesson-1-0-1, solved example-1-1-1)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(activity on lesson-1-0-3)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(activity on lesson-1-0-3, individual assignment-1-1-1)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(activity on lesson-1-1-2)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(additional resource-0-0-1)	*0*	*0*	*0*	*0*		*0*	*0*	*0*
(additional resource-0-0-2)	*1*	*1*	*1*	*1*		*1*	*0*	*0*
…								

FIGURE 4. Sample of the input matrix for homogeneous groups identification.

lecture-0-0-2, *lift* = 8.4). The value of the lift was very low in other academic years with the minimum in the last observed academic year.

E. SEASON 0 - OUTSIDE THE MAIN PERIOD

The records about stakeholders' activity after the period of final exams were also analyzed to find whether the stakeholders used to visit the course after they finished it as well as to identify preferred types of activities. The total count of stakeholders' sessions naturally decreased. However, since the proposed approach for the identification of frequent itemsets does not require any minimal count of the observed visits/sessions, this approach can also be used in this case. Again, the frequent itemsets with the *support* greater than 5% were considered.

On the other hand, it is essential to note that the qualitative conclusions cannot be generalized due to the limited number of sessions.

The common occurrence of the pairs of activity types was rare. Any type of activity was not dominant. The *lift* value was higher in comparison with the other observed seasons (except the academic years 13/14 and 15/16). This finding is in line with the expected behaviour of the stakeholders, who visited the course intending to find a particular study material or their solution of assignment.

As was mentioned earlier, the qualitative evaluation of the frequent 1-itemsets and 2-itemsets is influenced by the lower visit rate of the courses after their primary deployment during the academic year. Although there were identified several frequent 2-itemsets, their usefulness and potential generalization of the findings is limited.

VIII. ANALYSIS OF HOMOGENOUS GROUPS OF FOUND FREQUENT ITEMSETS

This section summarizes the individual steps of further analysis of the found frequent itemsets. As was already mentioned, the main aim of this analysis is to make comparison and identification of possible changes in stakeholder's behaviour or preferences in selecting particular activities easier. Therefore, a proposal, how to evaluate found frequent itemsets over different periods, is introduced in this section.

Therefore, the research question is, if the identified frequent itemsets with a sufficient minimum value of *support* can create any homogenous groups ($p > 0.05$) throughout the observed period and vice-versa if there are statistically significant differences between them ($p < 0.05$).

A. DATA PREPARATION

The first step of this second part of the research should begin again with the data preparation. The occurrence of the found frequent itemsets in individual seasons of the observed academic years was expressed as a matrix of unique values 0 or 1. Only the identified frequent 1-itemsets and 2-itemsets with the *support* greater than 5% were considered.

Subsequently, this matrix was further analyzed with the aim to find homogenous groups of frequent itemsets in the same seasons of the observed academic years, as well as during periods of the individual academic years.

B. COMPARISON OF SEASONALITY

There were not identified any statistically significant differences among the frequent items in the first observed season. The zero hypothesis was not rejected based on the Cochran Q test ($Q = 13.640$, $df = 7$, $p < 0.0580$) at the 0.05 significance level.

The zero hypothesis claims that the occurrence of the frequent itemsets is not depended on time. The table of multiple comparisons (Table 8) shows that only one homogeneous group ($p > 0.05$) was identified based on the average occurrence of found frequent itemsets of types of activities in the first identified season of observed academic years.

The zero hypothesis was rejected in the second observed season ($Q = 49.548$, $df = 7$, $p < 0.001$) (Table 9). Three homogenous groups were identified. The most frequent itemsets were identified in the academic year 10/11 (14%).

On the other hand, the lowest percentage of frequent itemsets was identified in the last two academic years, 16/17 and 17/18.

The statistically significant difference on the 0.05 significance level was proved between the academic year 16/17 and 14/15, as well as between 10/11 and academic years 11/12, 12/13, 13/14, 15/16, 16/17, 17/18.

TABLE 8. Homogenous groups of frequent itemsets of activities in the first season of the observed period of academic years.

year	incidence rate	1
17/18:1	5.87%	****
11/12:1	6.93%	****
15/16:1	8.00%	****
16/17:1	8.53%	****
14/15:1	9.60%	****
13/14:1	10.93%	****
10/11:1	11.20%	****
12/13:1	11.73%	****

TABLE 9. Homogenous groups of frequent itemsets of activity types in the second observed season.

year	incidence rate	1	2	3
16/17:2	3.47%	****		
17/18:2	5.87%	****	****	
15/16:2	6.13%	****	****	
12/13:2	6.13%	****	****	
13/14:2	7.20%	****	****	
11/12:2	7.20%	****	****	
14/15:2	10.67%		****	****
10/11:2	14.40%			****

TABLE 10. Homogenous groups of frequent itemsets of activity types in the third observed season.

year	incidence rate	1	2
15/16:3	2.40%		****
17/18:3	3.20%		****
16/17:3	3.47%		****
14/15:3	10.13%	****	
13/14:3	10.13%	****	
11/12:3	10.40%	****	
12/13:3	12.27%	****	
10/11:3	13.60%	****	

TABLE 11. Homogenous groups of frequent itemsets of activity types in the fourth observed season.

year	incidence rate	1
11/12:4	3.20%	****
17/18:4	3.20%	****
16/17:4	3.47%	****
14/15:4	3.73%	****
15/16:4	4.27%	****
13/14:4	4.53%	****
10/11:4	5.60%	****
12/13:4	6.67%	****

The analysis of the third season of observed academic years (Table 10) uncovered two homogenous groups ($Q = 79.124$, $df = 7$, $p < 0.001$). Simultaneously, this season was characterized by the decrease of the number of found frequent itemsets in the last three academic years. Statistically significant difference at the 0.05 level was proved between the academic years 10/11 to 14/15 and years 15/16 to 17/18.

TABLE 12. Homogenous groups of frequent itemsets of activity types in the fifth observed season.

year	incidence rate	1	2	3
15/16:0	2.67%	****		
13/14:0	3.20%	****		
11/12:0	3.47%	****		
10/11:0	4.00%	****	****	
16/17:0	4.27%	****	****	
17/18:0	6.67%	****	****	****
12/13:0	8.00%		****	****
14/15:0	10.40%			****

TABLE 13. Homogenous groups of frequent itemsets in the academic year 2010/2011.

period	incidence rate	1	2
10/11:0	4.00%		****
10/11:4	5.60%		****
10/11:1	11.20%	****	
10/11:3	13.60%	****	
10/11:2	14.40%	****	

TABLE 14. Homogenous groups of frequent itemsets in the academic year 2011/2012.

period	incidence rate	1	2	3
11/12:4	3.20%	****		
11/12:0	3.47%	****	****	
11/12:1	6.93%	****	****	****
11/12:2	7.20%		****	****
11/12:3	10.40%			****

Any statistically significant differences were not identified in the fourth observed season of the academic years ($Q = 10.733$, $df = 7$, $p < 0.1507$). However, the total share of frequent itemsets was very low (Table 11).

The last observed season of the academic years covers the period after the primary deployment of the course to the learning process. The zero hypothesis was rejected based on the Cochran Q test ($Q = 50.784$, $df = 7$, $p < 0.001$). The largest number of frequent itemsets was identified in the last third group, in the academic year 14/15.

Multiple comparisons based on the average incidence of the found frequent itemsets uncovered three homogenous groups (Table 12). A statistically significant difference between the groups at the 0.05 significance level was identified between the following academic years:

- 11/12, 13/14, 15/16 and years 12/13, 14/15,
- 10/11 a 14/15,
- 16/17 a 14/15.

C. COMPARISON OF DIFFERENT PERIODS IN ACADEMIC YEARS

The same approach was applied for the comparison of different periods in the individual academic years.

The zero hypothesis was rejected at the 0.05 significance level based on the results of the Cochran Q test showed for the academic year 10/11 ($Q = 53.385$, $df = 4$, $p < 0.001$). The

TABLE 15. **Homogenous groups of frequent itemsets in academic years 2012/2013, 2013/2014 and 2014/2015.**

period	incidence rate	1	2
12/13:2	6.13%	****	
12/13:4	6.67%	****	
12/13:0	8.00%	****	****
12/13:1	11.73%		****
12/13:3	12.27%		****

period	incidence rate	1	2
13/14:0	3.20%	****	
13/14:4	4.53%	****	
13/14:2	7.20%	****	****
13/14:3	10.13%		****
13/14:1	10.93%		****

period	incidence rate	1	2
14/15:4	3.73%		****
14/15:1	9.60%	****	
14/15:3	10.13%	****	
14/15:0	10.40%	****	
14/15:2	10.67%	****	

TABLE 16. **Homogenous group of frequent itemsets of activity types in academic years 2015/2016, 2016/2017, 2017/2018.**

period	incidence rate	1	2	3
15/16:3	2.40%	****		
15/16:0	2.67%	****		
15/16:4	4.27%	****	****	
15/16:2	6.13%		****	****
15/16:1	8.00%			****

period	incidence rate	1	2
16/17:3	3.47%	****	
16/17:2	3.47%	****	
16/17:4	3.47%	****	
16/17:0	4.27%	****	
16/17:1	8.53%		****

period	incidence rate	1	2
17/18:3	3.20%	****	
17/18:4	3.20%	****	
17/18:2	5.87%	****	****
17/18:1	5.87%	****	****
17/18:0	6.67%		****

largest number of frequent itemsets was identified in the second period of this year.

Multiple comparisons uncovered two homogenous groups based on the found frequent itemsets of individual types of activities (Table 13). Statistically significant difference at the 0.05 significance level was proven in the average number of occurrences of found frequent itemsets between period 0, 4 and 1, 2, 3.

Three homogenous groups were identified in 11/12 using the same approach ($Q = 33.638$, $df = 4$, $p < 0.001$). The statistically significant differences were found between periods 4 and 2, 3, as well as between periods 3 and 0 (Table 14).

Academic year 12/13 ($Q = 23.896$, $df = 4$, $p < 0.001$) can be characterized by two homogenous groups. A statistically significant difference at the 0.05 significance level was confirmed between the periods 2, 4 and 1, 3.

The statistically significant difference was found also in year 13/14 ($Q = 40.125$, $df = 4$, $p < 0.001$) between the periods 0, 4 and 1, 3.

Similarly, two homogeneous groups were identified in year 14/15 ($Q = 22.288$, $df = 4$, $p < 0.001$). A significant difference in the percentage of the found frequent itemsets is visible also in Table 15.

A significant decrease in the percentage of the found frequent itemsets in comparison with the previous years can be considered the main characteristic of the observed data in the last three years (15/16: $\mathrm{Q} = 28.071$, $\mathrm{df} = 4$, $\mathrm{p} < 0.001$; 16/17: $\mathrm{Q} = 26.784$, $\mathrm{df} = 4$, $\mathrm{p} < 0.001$; 17/18: $\mathrm{Q} = 13.263$, $\mathrm{df} = 4$, $\mathrm{p} < 0.05$) (Table 16). Three homogenous groups were found in the year 15/16 with the statistically significant difference between the seasons 0, 3 and 2, 1 as well as between season 1 and 4.

Two homogenous groups characterize the last academic years (16/17, 17/18). The statistically significant difference is visible between season 1 and other in academic year 16/17 as well as between the seasons 0 and 3,4 in the academic year 17/18.

D. PRACTICAL IMPLICATIONS OF IDENTIFIED HOMOGENEOUS GROUPS

The identified statistically significant differences between homogeneous groups of itemsets in observed periods could be considered the main contribution of this part of the research. At the same time, changing the percentual occurrence of the identified frequent itemsets provides other impulses for further qualitative interpretation of the research results and their impact on the management of the learning process.

For example, the comparison of homogeneous groups of individual seasons over academic years showed statistically significant changes in using the activities and resources in later academic years. Further detailed analysis of the possible reasons uncovered, that this change could be partially caused by the exchange of the teachers and the way, how they used course activities. While the first season remained unchanged over observed years, the third season showed a statistically significant difference between years 10/11 – 14/15 and 15/16 – 17/18.

On the other hand, comparison of identified homogeneous groups between the defined seasons of the individual years requires more precise analysis. The identified statistically significant differences vary between them. The frequent occurrence of the season with naturally low stakeholders' activity (4 or 0) in homogenous groups with other seasons suggests the stagnation or decrease in students' and teachers' external motivation to effectively utilize the opportunities, which the blended learning form, as well as the educational content of the course, provide.

IX. DISCUSSION AND CONCLUSION

The main contribution of the presented approach to the evaluation of frequent itemsets from the perspective of time can be reviewed from several points of view.

If the data pre-processing during a longer period is considered, the proposed approach emphasizes the requirement to implement one or more derived (abstract) variables, which allow transforming previously collected raw data of the log files to the more understandable form, which the domain expert can easier interpret.

Although this pre-processing phase is quite time-consuming and requires more data modification steps, the final transformation allows comparing the same types of items (in this case types of activities) over long period regardless of the fact, that the compared items could change their name or technical background. Moreover, the data was not aggregated too early in comparison to other similar approaches. Contrary, each row was extended with several time-based variables (*season*, *year*, *week*). As a result, the decision, if the observed variable creates an important feature of the research, was left for a later phase of the research.

The application of algorithms for frequent itemsets identification, which conforms defined threshold values of *support* measure showed that it is possible to uncover the changes in stakeholders' behaviour throughout the observed longer period of several academic years in the meaning of changes the stakeholders' preferences in choosing a particular type of activity as well as changes in the common use of more types of activities during one session.

The proposed approach allows estimating the probability of occurrence of a particular type of activity in the stakeholder's session.

The added value of the identified frequent itemsets is the ability to identify hidden pairs of activities, which were not considered important for achieving the aim required by the domain expert. However, the stakeholders already used them frequently together to solve partial tasks.

For example, a detailed analysis of the results showed that the students very often prefer to learn from solved examples or own previously solved individual assignments and bonus activities, which represent a source of knowledge for solving new tasks and assignments. As a result, the students do not use classical lectures as might be assumed considering their higher priority given by the domain expert.

A higher frequency of the bonus activity can be considered similarly interesting finding. It can be assumed that the students realize the advantage to participate in this kind of activity due to their contribution to the overall grading.

Consequently, the research was extended by the identification of the homogenous groups of activity types to understand better the identified changes in stakeholders' behaviour. The identification of statistically significant differences between the observed periods can be considered the main contribution of this part of the research. Similarly, a changing percentual representation of the identified frequent itemsets provides other impulses for further qualitative interpretation of the research results and their impact on the management of the learning process.

The proposed methodology can be generalized to other domains, for example, for identification of the seasonality and trends in requesting information on institutional websites or e-commerce [34].

The proposed methodology also has several limitations concerning the qualitative interpretation and usefulness of the found frequent itemsets.

The impact of the input dataset of logs should be mentioned in the first place. Its quality, completeness and size of dataset preordain the research results and their possible generalization. However, this obstacle was overcome, because the dataset used in the described research contained a larger number of unique users and their accesses to the VLE as used to be usual in the learning analytics research domain.

On the other hand, a large volume of data can decrease the overall performance of the presented approach. The presented technique is more laborious than others, mainly in the modeling phase. The reason is that it requires the same number of analysis as is the number of the observed seasons. Moreover, further analysis is realized after transforming the found rules to the input matrix for homogeneous groups identification.

The selection of the expert, who transformed selected attributes of the original dataset to new abstract (categorical) attributes, has the same importance. More objective approach, which could eliminate the possible subjective classification of the activities to the priority classes, can be based on the same approach as was used in [39]. The authors of this study asked a group of experts to assign values to the newly created attributes individually. Consequently, the authors statistically evaluated the measure of compliance between the experts. In this case, an incorrect assignment of the attributes (*bonus activities*, *activity on the lesson*) in academic year 11/12 could be revealed earlier.

The total count of the unique values, which new categorical attributes can acquire, can be considered the next limiting factor. Despite the fact the original attributes module and action were reduced, new attribute *type-detail* could acquire 61 different values. This large number could influence the final number of identified frequent itemsets and subsequently, the number/existence of the identified homogenous groups.

The definition of the threshold values of observed measures contributes to the overall interpretability and usefulness of the obtained results. As was mentioned in the related work section, the usefulness of the results can be improved using some other measures.

Finally, it is necessary to note that the decision into how many seasons the observed period should be divided can also influence the usefulness and interestingness of the found frequent itemsets. Again, the expert's voice plays a significant role in this task.

The approach presented in this paper contributes to the learning analytics research area focused on the pre-processing of logs stored in the VLE. Simultaneously, it provides an alternative approach to the research of behavioral changes of the VLE stakeholders throughout the longer period. As was mentioned earlier, this kind of research is relatively rare in the learning analytics domain [4]. Only a few researchers focused their research on the same topic [31],[33]. However, they applied other knowledge discovery tasks like time series analysis or data clustering based on different temporal characteristics.

The approach presented in this paper can be considered unique because it is not focused on the trend estimation or identification of the stakeholders with similar behaviour in the observed period. However, it is focused on the identification of statistically significant changes in the stakeholders' behaviour in the same part of several academic years.

Quantitative evaluation of the research results can be assessed from the course designer and teacher's point of view. The course designer can:

- focus on the types of activities preferred by the stakeholders,
- identify types of activities, which the stakeholders use more often together,
- review the role of additional, optional activities and study resources,
- increase the visit rate of the course using conditional activities and completion of individual activities in the dependence on observed periods,
- re-design the structure and the content with respect to the decrease in the probability of occurrence of the particular type of activity in the individual sessions,
- adapt the portfolio of provided types of activities to the different periods considering the observed behaviour of the stakeholders.

The teacher or mentor, who repeatedly opens the course, can be regularly informed about the changes in the structure of the most visited types of activities in the observed period of several years. However, the probability of the accesses of the identified frequent itemsets of activity types and their correlation is considered more interesting, because the teacher can see how the students learn throughout several academic years. These observations can help the teacher to combine different types of activities more suitably to engage the students to be more active in the e-learning course environment.

The future research direction can be focused on the evaluation of the impact of other measures of frequent pattern mining on the usefulness of the obtained results and their subsequent generalization. Simultaneously, the research will be extended to identify other typical periods or seasons of the observed academic years, compare and map the results with other research initiatives, which try to characterize the relation between the learning analytics and learning design, as well as to identify different learning styles of stakeholders based on the differences in their behaviour throughout the observed period. Finally, the proposed methodology can be evaluated using the standardized dataset [47] to confirm the usefulness of the proposed approach and its wider use in the learning analytics domain.

REFERENCES

[1] F. Chen and Y. Cui, "Utilizing student time series behaviour in learning management systems for early prediction of course performance," *J. Learn. Anal.*, vol. 7, no. 2, pp. 1–17, Sep. 2020.

[2] J. Saint, D. Gašević, W. Matcha, N. A. Uzir, and A. Pardo, "Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning," in *Proc. 10th Int. Conf. Learn. Anal. Knowl.*, Mar. 2020, pp. 402–411.

[3] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le, "A survey of itemset mining," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 7, no. 4, p. e1207, Jul. 2017.

[4] M. Drlik and M. Munk, "Understanding time-based trends in stakeholders' choice of learning activity type using predictive models," *IEEE Access*, vol. 7, pp. 3106–3121, 2019.

[5] E. García, C. Romero, S. Ventura, C. D. Castro, and T. Calders, "Association rule mining in learning management systems," in *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, Eds. London, U.K.: Chapman & Hall, 2011, pp. 93–106.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. R. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," Tech. Rep., 2000.

[7] J. M. Luna, P. Fournier-Viger, and S. Ventura, "Frequent itemset mining: A 25 years review," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 9, no. 6, Nov. 2019, Art. no. e1329.

[8] J. M. Luna, P. Fournier-Viger, and S. Ventura, "Extracting user-centric knowledge on two different spaces: Concepts and records," *IEEE Access*, vol. 8, pp. 134782–134799, 2020.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. SIGMOD Rec.*, vol. 22, Jun. 1993, pp. 207–216.

[10] H. Toivonen, "Frequent itemset," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2010, p. 418.

[11] H. Toivonen, "Frequent pattern," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2010, pp. 418–422.

[12] C. C. Aggarwal and J. Han, *Frequent Pattern Mining*. Cham, Switzerland: Springer, 2014.

[13] S. Ventura and J. Luna, *Supervised Descriptive Pattern Mining*. Cham, Switzerland: Springer, 2018.

[14] J. Skalka, "Data processing methods in the development of the microlearning-based framework for teaching programming languages," in *Proc. 12th Int. Sci. Conf. Distance Learn. Appl. Inform. (DIVAI)*, M. Turcani, Z. Balogh, M. Munk, J. Kapusta, and L. Benko, Eds., 2018, pp. 503–512.

[15] E. García, C. Romero, S. Ventura, and C. D. Castro, "An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering," *User Model. User-Adapted Interact.*, vol. 19, nos. 1–2, pp. 99–132, Feb. 2009.

[16] D. L. Bazaldua, R. S. J. D. Baker, and M. O. S. Pedro, "Comparing expert and metric-based assessments of association rule interestingness," in *Proc. 7th Int. Conf. Educ. Data Mining*, J. Stamper, Z.Pardos, M. Mavrikis, and B. M. McLaren, Eds. London, U.K.: Institute of Education, Jul./Jul. 2014, pp. 44–51.

[17] B. Daniel, "Big data and analytics in higher education: Opportunities and challenges," *Brit. J. Educ. Technol.*, vol. 46, no. 5, pp. 904–920, Sep. 2015.

[18] B. K. Daniel, *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*. Cham, Switzerland: Springer, 2017.

[19] A. Merceron and K. Yacef, "Measuring correlation of strong symmetric association rules in educational data," in *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, Eds. London, U.K.: Chapman & Hall, 2011, pp. 245–256.

[20] J. M. Luna, C. Romero, J. R. Romero, and S. Ventura, "An evolutionary algorithm for the discovery of rare class association rules in learning management systems," *Int. J. Speech Technol.*, vol. 42, no. 3, pp. 501–513, Apr. 2015.

[21] Y.-M. Huang, J.-N. Chen, and S.-C. Cheng, "A method of cross-level frequent pattern mining for Web-based instruction," *J. Educ. Technol. Soc.*, vol. 10, pp. 305–319, Jul. 2007.

[22] N. A. Uzir, D. Gašević, J. Jovanović, W. Matcha, L.-A. Lim, and A. Fudge, "Analytics of time management and learning strategies for effective online learning in blended environments," in *Proc. 10th Int. Conf. Learn. Anal. Knowl.*, Mar. 2020, pp. 392–401.

[23] Q. Nguyen, "Rethinking time-on-task estimation with outlier detection accounting for individual, time, and task differences," in *Proc. 10th Int. Conf. Learn. Anal. Knowl.*, 2020, pp. 376–381.

[24] B. Saleh and F. Masseglia, "Discovering frequent behaviors: Time is an essential element of the context," *Knowl. Inf. Syst.*, vol. 28, no. 2, pp. 311–331, Aug. 2011.

[25] J. M. Ale and G. H. Rossi, "An approach to discovering temporal association rules," presented at the ACM Symp. Appl. Comput., vol. 1, Como, Italy, 2000.

[26] T. Xie, Q. Zheng, and W. Zhang, "Mining temporal characteristics of behaviors from interval events in e-learning," *Inf. Sci.*, vol. 447, pp. 169–185, Jun. 2018.

[27] T.-Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, "Behavior-based grade prediction for MOOCs via time series neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 5, pp. 716–728, Aug. 2017.

[28] J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, "Identifying at-risk students for early interventions—A time-series clustering approach," *IEEE Trans. Emerg. Topics Comput.*, vol. 5, no. 1, pp. 45–55, Jan./Mar. 2017.

[29] C. Herodotou, B. Rienties, M. Hlosta, A. Boroowa, C. Mangafa, and Z. Zdrahal, "The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study," *Internet Higher Educ.*, vol. 45, Apr. 2020, Art. no. 100725.

[30] M. J. Mahzoon, M. L. Maher, O. Eltayeby, W. Dou, and K. Grace, "A sequence data model for analyzing temporal patterns of student data," *J. Learn. Anal.*, vol. 5, no. 1, pp. 55–74, Apr. 2018.

[31] M. S. Boroujeni and P. Dillenbourg, "Discovery and temporal analysis of MOOC study patterns," *J. Learn. Anal.*, vol. 6, no. 1, pp. 16–33, Apr. 2019.

[32] N. Quan, B. Rienties, and L. Toetenel, "Unravelling the dynamics of instructional practice: A longitudinal study on learning design and VLE activities," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, 2017, pp. 168–177.

[33] Q. Nguyen, M. Huptych, and B. Rienties, "Using temporal analytics to detect inconsistencies between learning design and students' behaviours," *J. Learn. Anal.*, vol. 5, no. 3, pp. 120–135, Dec. 2018.

[34] M. Munk, A. Pilkova, L. Benko, and P. Blažeková, "Pillar 3: Market discipline of the key stakeholders in CEE commercial bank and turbulent times," *J. Bus. Econ. Manage.*, vol. 18, no. 5, pp. 954–973, Oct. 2017.

[35] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Hum. Behav.*, vol. 89, pp. 98–110, Dec. 2018.

[36] W. Han, J. Borges, P. Neumayer, Y. Ding, T. Riedel, and M. Beigl, "Interestingness classification of association rules for master data," in *Proc. Ind. Conf. Data Mining*, Cham, Switzerland, 2017, pp. 237–245.

[37] B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing the subjective interestingness of association rules," *IEEE Intell. Syst.*, vol. 15, no. 5, pp. 47–55, Sep. 2000.

[38] A. Merceron and K. Yacef, "Interestingness measures for association rules in educational data," in *Proc. 1st Int. Conf. Educ. Data Mining*, R. S. J. D. Baker, T. Barnes, and J. E. Beck, Eds. Montreal, QC, Canada, Jun. 2008, pp. 57–66.

[39] M. Munk, M. Drlik, L. Benko, and J. Reichel, "Quantitative and qualitative evaluation of sequence patterns found by application of different educational data preprocessing techniques," *IEEE Access*, vol. 5, pp. 8989–9004, 2017.

[40] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2 ed. New York, NY, USA: Springer, 2011.

[41] M. Drlik, J. Skalka, P. Svec, and J. Kapusta, "Proposal of learning analytics architecture integration into university IT infrastructure," in *Proc. IEEE 12th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2018, pp. 265–270.

[42] J. Obonya and J. Kapusta, "Identification of important activities for teaching programming languages by decision trees," in *Proc. 12th Int. Sci. Conf. Distance Learn. Appl. Inform. (DIVAI)*, M. Turcani, Z. Balogh, M. Munk, J. Kapusta, and L. Benko, Eds., 2018, pp. 481–490.

[43] 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, pp. 27–36, Feb. 2013.

[44] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009.

[45] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," presented at the 20th Int. Conf. Very Large Data Bases, 1994.

[46] M. Hahsler, B. Grün, and K. Hornik, "Arules—A computational environment for mining association rules and frequent item sets," *J. Stat. Softw.*, vol. 14, no. 15, p. 25, 2005.

[47] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Sci. Data* vol. 4, Nov. 2017, Art. no. 170171.

MARTIN DRLIK 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 research study program of the Theory of Computer Science Education from Constantine the Philosopher University in Nitra, Slovakia, in 2009. Since 2019, he has been an Associate Professor with the Department of Informatics, Constantine the Philosopher University in Nitra. He is currently a member of the Knowledge Discovery Research Group. His research interests include learning analytics, educational data mining, software engineering, and database systems.

MICHAL MUNK received the M.S. degree in mathematics and informatics and the Ph.D. degree in mathematics from Constantine the Philosopher University in Nitra, Slovakia, in 2003 and 2007, respectively. In 2018, he was appointed as a Professor in system engineering and informatics with the Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic. He is currently a Professor with the Department of Informatics, Constantine the Philosopher University in Nitra. He is the also Head of the Knowledge Discovery Research Group. His research interests include data analysis, Web mining, and natural language processing.

JAN SKALKA received the M.S. degree in mathematics and informatics with the Department of Informatics, Constantine the Philosopher University in Nitra, Slovakia. He defended his Ph.D. thesis in 2004 in the research study program of Technology of Education from Constantine the Philosopher University in Nitra, where he has been the Head of the Department of Informatics, since 2018. He is currently a member of the Knowledge Discovery Research Group. His research interests include information systems implementation and integration, blended learning and e-learning application in education, programming learning and teaching, and development of applications to support education.