The Four Seasons: Identification of Seasonal Effects in LMS usage data

Monika Andergassen
Information Systems and New Media
Vienna University of Economics and Business (WU)
Vienna, Austria
monika.andergassen@wu.ac.at

Gustaf Neumann
Information Systems and New Media
Vienna University of Economics and Business (WU)
Vienna, Austria
gustaf.neumann@wu.ac.at

Felix Mödritscher
Information Systems and New Media
Vienna University of Economics and Business (WU)
Vienna, Austria
felix.moedritscher@wu.ac.at

ABSTRACT[1]

Learning Management Systems (LMSs) are widely used by organizations to provide and manage educational activities. Particularly in higher education, the application of LMS platforms is well documented and evaluated in the literature for at least one decade, whereby evaluation is often restricted to user-oriented analysis of the acceptance, usefulness and usability and rarely relies on real data-sets. Previous research revealed that the usage patterns of web users and mobile users highly depend on the time period within a semester. Therefore, this paper specifically addresses the question how to identify and compare seasonal effects on the basis of an anonymized data-set. After proposing an Educational Data Mining based method for analyzing log files of LMS platforms and elaborating related work, we report a case study in which we compare the usage behavior of four different seasons. It shows that not only the intensity of platform usage but also certain activities of LMS users are highly dependent on the season. Moreover, seasons can be characterized e.g. through rank/frequency plots of *n*-grams or principal components of the browsing sessions in the period of time. The paper provides evidence that the detection of seasonal effects can be used for improving the navigation structures and personalization of LMS systems.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications: *Data mining*, I.5.1 [**Computing Methodologies**]: Pattern Recognition: *Models – Structural*, J.1 [**Computer Applications**]: Administrative Data Processing: *Education*.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Learning Management Systems, Educational Data Mining, Browsing Sessions, Principal Component Analysis, Seasonal Effects, Learning Analytics.

1. INTRODUCTION

From the perspective of organizations, Learning Management Systems (LMSs) are a key technology for providing and managing learning and related resources in various application scenarios, such as higher or further education and workplace learning [19]. At the same time, the increasing penetration of mobile devices into society [6] requires the providers of LMS platforms to adapt their systems to a variety of end-user devices. Particular challenges include limited bandwidth, different resolutions and input devices, where the adaptation of the user interface and the content can ease the interaction of end-users and improve the responsiveness of the device.

Concerning the personalization of educational technology towards learners, much attention is paid to the field of Learning Analytics and thus also to "*the interpretation of a wide range of data produced by and gathered on behalf of students in order to assess academic progress, predict future performance, and spot potential issues*" [11]. From the perspective of organizations, the analysis of LMS usage behavior relates also to Academic Analytics which also focuses on applying statistical techniques and predictive models in order to help institutions to fulfill their academic missions [3].

2. RELATED WORK

In former research on Learning and Academic Analytics (cf. [16]) we compared the use of the mobile version of an LMS (i.e. an access point which is optimized for mobile devices, such as smartphones and tablets) to the use of the full version of the LMS (i.e. a site which is built for being accessed with standard web browsers). It was shown that mobile users, i.e. users accessing the LMS through smartphones, tablets and so forth, tend to have shorter browsing sessions and rather look up information while web users have longer sessions which include more course-specific activities such as using forums and solving multiple-choice questions. The clustering of similar sessions resulted in some big clusters such as '*exam review-logout*'. However, these observations are restricted to the LMS log file of one specific day, while findings also indicated that usage behavior strongly fluctuates through different seasons within an academic year.

A temporal analysis of student behavior in online courses based on LMS log files is also done, for instance, in [9] and [13]. In [9], the persistence of students in a course throughout a semester is analyzed through the cumulative overall activities of individual students. The analysis results in the identification of five types of persistence: '*low-extent users*', '*late users*', '*online quitters*', '*accelerating users*' and '*decelerating users*'. In [13], the average duration and the average intervals between student user sessions in discussion forums are clustered and lead to the identification of different learner types such as '*committed*', '*directed*' and '*strategic*'. Both studies concentrate on the overall number of activities of students and do not research different types of activities or sessions undertaken by the students.

Temporal analyses with focus on action types are conducted in [1] and [14]. In [14], learning behaviors are identified and contextualized by performance evolution between groups of

[1] Presented at the DAILE'13: Workshop on Data Analysis and Interpretation for Learning Environments, 28 January 2013 - 01 February 2013,Villard-de-Lans, France

students. Methodologically, user activities are abstracted from LMS log files, categorized into five primary categories such as '*READ*', '*LINK*' and '*QUER'*, and provided with additional metrics such as whether a source is read for the first time or repeatedly. Through a combination of sequence mining techniques, action sequences are then identified and their frequencies in the user sessions are calculated. Four distinct categories of frequent patterns emerge which are then compared to the level of performance of students over time. Findings include that high-performing students have a different reading behavior than low-performing students, for instance by re-reading pages more often. Similarly, students' characteristic learning behaviors in Intelligent Tutoring Systems (ITS), in particular their self-regulated-learning, are investigated in [1]. Again, user activities abstracted from ITS log files are categorized using the three categories '*Reading*', '*Monitoring*' and '*Strategy*'. Then, three clusters of students are identified using Expectation Maximization, and characterized by different prior knowledge of the topic, learning performance, and strategies. Finally, typical activity patterns for students of the various clusters are identified using sequence mining techniques.

All these studies concentrate on the learning activities of the individual students and investigate their temporal development either in terms of persistence or in terms of performance. The students are then categorized according to their learning activities [1, 9, 13] or their sequences of learning activities [1, 9]. However, our findings indicate that learning activities and user sessions might not only differ on the level of the individual student, but also on the level of different seasons [16].

Consequently, the focus of the actual paper lays on the identification of differences in the learning activities between various seasons throughout a semester. In particular, we address the following research questions:

1. How can students' learning activities and user sessions in an LMS be identified and compared?
2. Do the learning activities and user sessions depend on seasonal effects such as exam periods and holidays?

Section 3 briefly describes our approach which is based on web log analysis (WLA) and educational data mining (EDM). After characterizing the available data-set, the data mining process for identifying seasonal effects is depicted in detail. Section 4 reports on a case study on the Learn@WU platform, the LMS which is institutionally offered by the Vienna University of Economics and Business (WU). The previously described analysis approach is applied to identify seasonal effects of learning activities within a semester. Finally, Section 5 discusses the findings and concludes the paper.

3. ANALYSIS OF LMS LOG DATA AND DETECTION OF USAGE PATTERNS

Starting with considerations from the field of Web Analytics [7], usage data can be gathered on the server side (e.g., web server logs), on the client side (e.g., page tagging) or through hybrid methods.

The approach presented in this paper fully relies on the analysis of web server logs and thus could be applied in general to any web-based LMS. Table 1 shows the composition of a typical log file entry. Slightly extending the common log format [10], the advantage of LMS data-sets is that it can be extended with parameters such as unique user identifiers for each request, as LMS platforms are closed systems which require authentication.

Table 1. Composition of a log file entry, including an LMS-specific field (marked with *)

IP-Address of Client (anonymized)
12.34.56.78
Remote User
-
HTTP-User
-
Timestamp
[09/Apr/2012:00:06:11 +0200]
HTTP-Request
GET /dotlrn/?pnum=2&pname=news_portlet HTTP/1.0
HTTP-Response
200
HTTP-Response Size
15009
HTTP-Referrer
https://learn.wu.ac.at/dotlrn/?pnum=6&pname=tlf_homework_portlet
User Agent
Mozilla/5.0 (iPhone; CPU iPhone OS 5_1 like Mac OS X) AppleWebKit/534.46 (KHTML, like Gecko) Version/5.1 Mobile/9B176 Safari/7534.48.3
Anonymized User Identifier *
12345678

We processed the log files following the three phases of Web Usage Mining [21], namely preprocessing, pattern discovery and pattern analysis. Figure 1 visualizes the educational data mining process for identifying seasonal effects in LMS usage data.

During **usage preprocessing**, sensitive data, like the user identifiers and the IP addresses, was anonymized using a simple k-anonymization [5] technique in a first step. Then, the URLs were simplified, and the requests were grouped into browsing sessions (i.e. on the basis of the anonymized user identifier, the IP addresses and the timeout threshold). Hereby, the most notable advantage of closed systems like LMSs is that the user identifier can be utilized for reconstructing the browsing sessions, which solves two big challenges of Web Usage Mining [21]: (a) one user accessing the LMS through various devices and IP addresses, and (b) different users accessing it from the same IP address, for instance with different browsers on the same computer. Requests of anonymous users (i.e. requests before authentication or by web robots) were being filtered out of the data-set. Finally, some users might access the LMS various times in one day. In this latter case, we defined the end of a session through either an active log out by

the user or by the passing of more than 20 minutes before the next activity.

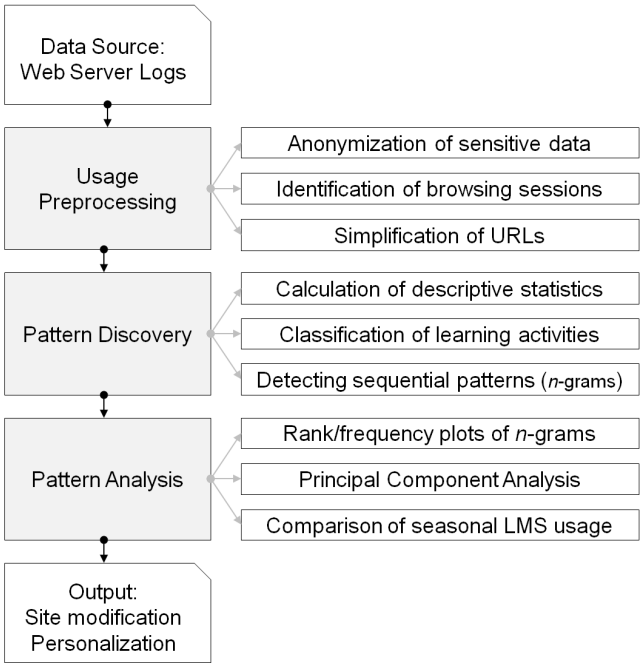

Figure 1: Log file analysis process to identify seasonal effects (based on the Web Usage Mining process [21])

To prepare the data for pattern discovery and analysis, some further preprocessing was done in the first step. Derived requests (requests for embedded content such as images, JavaScript and CSS) were filtered and discarded.

For **pattern discovery**, we developed a naming scheme to deduce learning activities from session entries. In contrast to behavior or interest mining from log files [22], we focused on classifying the session activities of the users as learning activities. With respect to supervised inductive learning [23] we used a training set for implementing a classifier that maps request information to learning activities. In particular, the classification is based on the URL, the HTTP method, the query parameter and the time spent on the page. A more detailed analysis might require including additional information in the classification input, such as post data or detailed information about the interaction items from the database. According to the applications and services available to the students on our LMS, we named the activities based on the objects and operations, on which these were applied. Examples of such objects are '*Wiki*', '*Forum*', '*Calendar*', etc. We treated POST-requests as a different activity as e.g. GET-request; therefore as writing of a forum post ('*Forum-post*'), solving of a multiple-choice question ('*Excs-problem*'), and editing a wiki page ('*Wiki-write*'). Furthermore, we distinguished operations depending on the time a user would spend on them. Assuming that a user can not consciously read or learn a content if he spends only a minimum of time on it, we tagged operations lasting less than 10 seconds with '*look*' (or '*memorize*', or '*navigate*' depending on the context) and activities lasting longer than 10 seconds with '*read*' (or '*view*' in the case of multiple-choice questions). Finally, we provided distinct names for activities relating to navigating on the portal page. The portal page aggregates new information such as new learning materials, new forum posts, new homework tasks, etc., of all courses a student is subscribed to.

This resulted in a non-exhaustive naming scheme for **79 learning activities**, which one could modify according to the emphasis of the individual research. The naming scheme is non-exhaustive because the methods described above do not tackle all the possible request URLs of the platform. The methods pointed at capturing the most widely used activities. Rarely occurring activities such as, for instance, the customization of the personal portal page, are not yet captured by the list of activities. Future work should aim at complementing the existing list with the missing activities. A list of all the 79 types of learning activities so far identified for this study can be found in Appendix A. Examples of this naming scheme include:

- *Personal-Portal-Read:* Reading the personal portal page. The personal portal page is the first page a student gets after login and the point where all personal information such as course subscriptions, deadlines and news are displayed
- *Community-Portal-Read:* Reading the community portal page. Each course has its community portal page were all course information is aggregated.
- *Excs-Problem-View:* Solving a multiple choice question, where the tag *'view*' indicates that the student spent more than 10 seconds on the page.
- *Excs-Score-View:* Watching the score and right solution of the multiple choice question just solved, where the tag '*view*' indicates that the student spent more than 10 seconds on the page.
- *Excs-Score-Memorize:* Watching the score and right solution of the multiple choice question just solved, where the tag '*memorize*' indicates that the student spent less than 10 seconds on the page.
- *Forum-Look:* Watching a forum thread for less than 10 seconds.
- *Forum-Read:* Watching a forum thread for more than 10 seconds.
- *Forum-Post:* Writing a forum post.

The naming scheme can be applied to all sessions extracted from a log file. An example of a browsing session is represented in Table 2. In this session with the ID 1003, the user starts from the personal portal page (*Personal-Portal-Read*), then continues to the portal page of a course (*Community-Portal-Read*), then shortly watches the forum (*Forum-Look*) und consecutively, reads a forum post in more detail (*Forum-Read*).

Table 2: Example of a session with naming scheme for learning activities

Action ID	*Learning activity*	*Duration*
1	Personal-Portal-Read	19
2	Community-Portal-Read	10
3	Forum-Look	9
4	Forum-Read	18
5	Search-Navigate	41
6	Search-Navigate	40
7	Search-Navigate	9
8	Personal-Portal-Read	n.a.
Summary: session 1003; duration 146; activities 8; mobile 1		

Then, he uses the LMS search function (*Search-Navigate*) three times, perhaps he does not find immediately what he is searching for. The session concludes with returning to the personal portal page (*Personal-Portal-Read*). The duration of each activity except from the last activity in a session is calculated as the timespan between two server requests logging two activities. The duration of the last avitivty in each session cannot be determined, since the system has no information about the time the user spent on this page.

Also, some descriptive statistical data, such as the number and length of sessions, the number of unique users, the requests per user and session, and the frequencies of specific learning activities were collected.

A first method for **pattern analysis and the identification of seasonal effects** was to compare the statistics of data-sets retrieved from different seasons (period of time, e.g. weeks) with each other. Moreover, the frequency distribution of the 10 most volatile learning activities was plotted over the 4 weeks.

Furthermore, *n*-gram analysis was applied to activities of the analyzed sessions; in particular 1-gram, 2-gram and 3-gram analyses were used to detect frequencies of learning activities and sequences of adjacent learning activities in the sessions. As part of pattern analysis, we examined the rank-frequency distribution of 1-grams (the single learning activities) and 3-grams (sequences of three learning activities) in general and per season.

Finally, we applied Principal Component Analysis (PCA) for identifying patterns in data of high dimension. With high dimension we referr to the browsing sessions which can be characterized by a vector of learning activities a user performed within the session. We compared the principal components of the data of the seasons.

Since our approach presented above is fully based on web server logs, it can be applied to a multiplicity of learning management systems and is not limited to a specific setting. However, as a distinct example of use, in the next section we present the analysis of learning activities of the LMS platform Learn@WU.

4. COMPARISON OF SEASONAL EFFECTS OVER DIFFERENT PHASES IN A SEMESTER

The LMS Learn@WU which is in use at the Vienna University of Economics and Business (WU), provided the data for the case study described in this section. Learn@WU has been in use for 10 years at the WU and is one of the most intensely used e-learning platforms world-wide. Up to 2,500 concurrent users solve up to 600,000 interactive exercises per day [16], leading to peaks of 3.8 million page views per day.

At the WU, Learn@WU enjoys a high acceptance. All courses offered by the university are mapped in this LMS. Thus, every lecturer has the possibility to offer the whole range of e-learning applications in their classes. At present, the LMS is used for a variety of blended learning scenarios but decidedly not for distance learning courses. While the broadest field of application comprises the beginning phase of the Bachelor programs, Learn@WU is also in use in the advanced courses of the Bachelor program, as well as in the Master and PhD programs.

Since May 2011, Learn@WU has been available in a version optimized for mobile devices, in particular smartphones. This '*mobile*' version offers all functionalities of the '*full*' version, except that the user interface is optimized for touch screens and for small screen sizes. Additionally, the number and size of files to display the user interface (e.g., JavaScript and CSS-files) were optimized to reduce the amount of server requests and the page file sizes. This mobile version provided the data for the analysis presented in this paper.

4.1 Method and data-set for this case study

We compared log files of four distinct weeks of the summer semester 2012 which we estimated as characteristic weeks in the WU semester plan. In total, we processed 264,837 log file entries. The first week comprised the Easter holiday, a period with usually modest traffic on Learn@WU (named '*holiday*'). The second week comprised the week before the mid-term exam period (named '*pre-exam*'). Students usually learn intensely for their exams in this week. The third week comprised the mentioned mid-term exam week (named '*exam*'). At the WU, there are 6 exam weeks distributed among the academic year, and we estimated that the students' learning activities on Learn@WU before and during the exam weeks might differ from other periods. Finally, the fourth week comprised the week after the mid-term exam week, and thus a week with potentially somewhat usual student activities on Learn@WU (named '*post-exam*').

To prepare the data, we extracted 3 log files of each of the 4 weeks and calculated the average percent values of all learning activities within each week. This measure was taken to avoid potential outliers which might occur on a single day.

4.2 Statistical indicators of the four seasons

Table 3 gives an overview of selected statistical indicators (i.e. typical Web Analytics metrics) which were calculated in the pattern discovery phase. It appears that the overall number of requests multiplies almost tenfold between the holiday week (n=8,512) and the exam week (n=79,958). The increase of the overall sessions is even bigger, leading to almost 12 times the overall sessions in the exam week (n=11,190) compared to the holiday week (n=960).

Table 3: Comparison of Web Analytics metrics of the 4 weeks

	Week 1	Week 2	Week 3	Week 4
No. (page) requests	8512	36689	79958	33686
Avg. request duration [sec]	41,24	41,00	45,47	40,37
No. sessions	960	4226	11190	5772
No. visitors (users)	231	838	1755	1213
No. sessions per user	4,16	5,04	6,38	4,76
Avg. no. requests/session	8,87	8,68	7,15	5,84
Avg. session duration [sec]	366,12	356,70	325,34	235,90
No. unique activities	107	113	120	114
Avg. activity frequency	79,55	324,68	666,32	295,49
Activity frequency: st.dev.	169,48	797,08	1892,32	825,07

The average interaction frequency, and thus the mean number of individual learning activities increases strongly between the holiday week (n=79.55) and the exam week (n=666.32). At the same time the standard deviation also strongly increases (169.48 compared to 1892.32), indicating that the number of occurrences of learning activities is highly volatile in the learning periods before and during the exam week. The next section will expand on this observation.

4.3 *N*-gram analysis

The get a clearer indication about which kind of interaction sequences users are likely to perform in different seasons, we conducted *n*-gram analysis on the learning activities. *N*-gram

analysis is used, for instance, in statistical natural language processing to predict the probability by which a word appears after another [18]. A sequence of words, for instance the sentence '*We conducted n-gram analysis*' can be split up in chunks of n adjacent words. This would lead, in the above example, to four 1-grams or unigrams (if n=1), three 2-grams or bigrams (if n=2), two 3-grams or trigrams (if n=3) and one 4-gram (if n=4). An example can be found in [20], where all the articles published in the '*Communications of the ACM*' over 10 years were analyzed using *n*-grams, and in [8] with Google search queries.

In our research, we adopted *n*-gram analysis to analyze adjacent learning activities. We developed 1-grams and 3-grams of the learning activities of all 4 seasons. We compared them over the 4 seasons by generating rank/frequency plots [17] and by calculating the volatility of the *n*-grams.

Figure 2 shows a rank/frequency plot of the **1-gram learning activities** in the holiday week (week 1) of the investigation. Table 4 lists the data of all four seasons.

The learning activities are sorted in decreasing order of appearance on the x-axis of Figure 2. Their frequencies are represented on the y-axis. It appears that a few learning activities have high frequencies and many learning activities have only low frequencies. As power law distributions were observed e.g. for items which are shared and used in (music) communities [4], we calculated the parameters of such distributions (power law slope α, goodness of fit R^2) according to the maximum likelihood with the 'igraph' package of the R system (cf. http://cran.r-project.org).

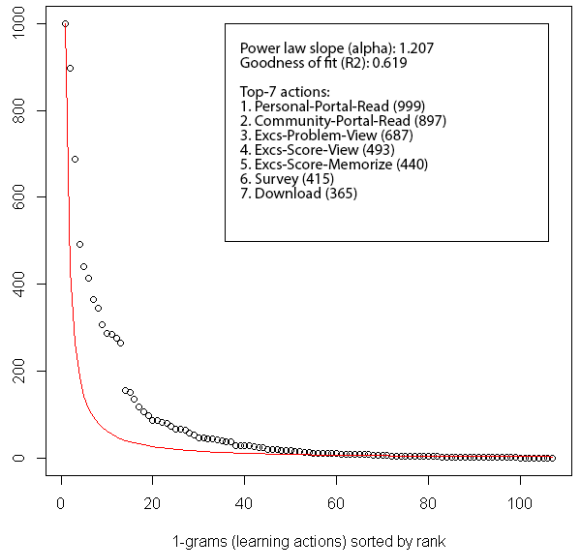

Figure 2: Rank/frequency plot and power law approximations of 1-gram learning activities in holiday week (week 1)

The 1-gram rank/frequency plot shows high frequencies of the learning activities *'Personal-Portal-read*' and '*Community-Portal-read*', and for viewing and solving multiple choice exercises. Yet, the LMS was used more intensely in week 3, leading to higher numbers of frequencies in week 3. Moreover, the goodness of fit of the power law approximation is clearly higher in week 3 (61.9% vs. 89.5%), as is shown in Table 4. Another observation deals with the activity '*survey'* (short for '*Survey-answer*'). The survey was only conducted in the holiday season (week 1), so these items do not appear in the *n*-grams of the other weeks.

Table 4: Comparison of 1-grams between 4 seasons

	Week 1	Week 2	Week 3	Week 4
R^2	0,619	0,622	0,895	0,807
alpha	1,208	1,076	1,028	1,087
Top 10 1-grams	Pers-Port-Read	Excs-Problem-View	Exam-Result	Pers-Port-Read
	Comm-Port-Read	Pers-Port-Read	Pers-Port-Read	Comm-Port-Read
	Excs-Prob-View	Excs-Score-View	Excs-Score-View	Exam-Result
	Excs-Score-View	Comm-Port-Read	Comm-Port-Read	Forum-Read
	Excs-Score-Memo	Excs-Score-Memo	Navigate	Forum-Look
	Survey	Download	Forum-Read	Navigate
	Download	File	Sample-exam	File
	Navigate	Navigate	Forum-Look	Download
	Excs	Excs	Excs-Prob-View	Course-Info
	Excs-Detail	Forum-Read	Excs-Score-Memo	Pers-Port-Calendar

As a next step, we calculated the mean values and standard deviations of all learning activities over the 4 weeks in order to detect the most fluctuating activities. The frequency distribution of these volatile learning activities is presented in Figure 3.

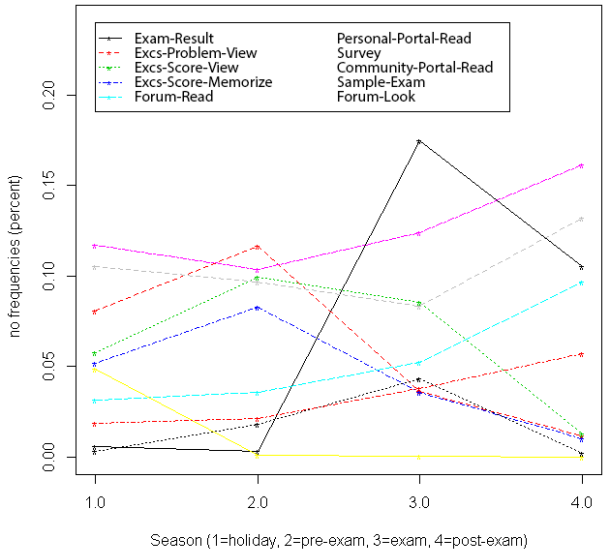

Figure 3: Frequency distribution of 10 most fluctuating learning activities (1-grams) over 4 measurement periods

The learning activity *'Exam-Result*' shows the strongest standard deviation over the 4 weeks. This activity, where students view their grades in Learn@WU, is obviously important to students only after exams and therefore little used before the exams.

The learning activity '*Excs-Problem-view*' which is related to viewing a multiple-choice question has a high frequency in the pre-exam week and drops afterwards. This can be explained through the fact that students prepare for their exams through solving multiple-choice questions. The learning activities '*Excs-Score-view*' and '*Excs-Score-memorize*' both show drops in the exam week for similar reasons than explained just above. However, it is interesting to note that *memorizing* multiple-choice answers drops more than *viewing* multiple-choice answers. This might be an indication that students memorize less and actively learn more shortly before their exam.

Similarly, the **3-gram learning activities** show frequencies which come close to a power-law distribution (See Figure 4 for the 3-grams of the holiday week).

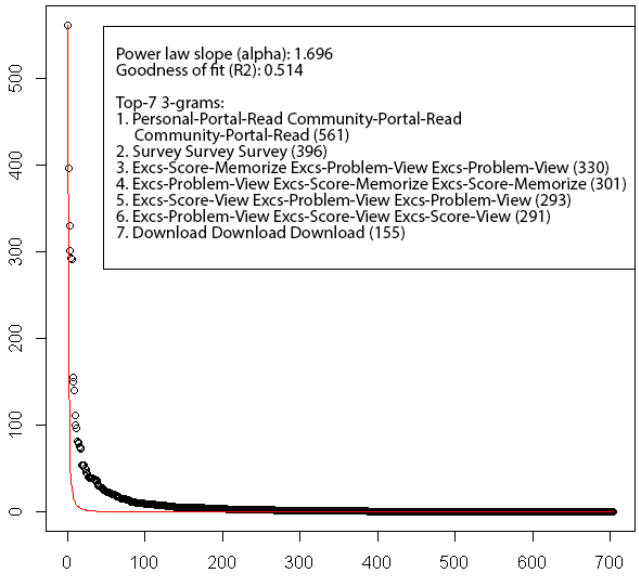

Figure 4: Rank/frequency plot and power law approximations of 3-gram learning activities in holiday week (week 1)

A combination of visits of the personal portal page and the community portal pages (*Personal-Portal-read Community-Portal-read Community-Portal-read*) is prevalent in all the investigated seasons, as can be seen in Table 5. This indicates that these pages are important reference points for students and that through these pages they might find much information which is typically searched with a mobile device.

Table 5: Comparison of 3-grams between 4 seasons

Week 1	
R^2	0,514
alpha	1,696
Top 3 3-grams	Pers-Portal-Read Comm-Portal-Read Comm-Portal-Read
	Survey Survey Survey
	Excs-Score-Memo Excs-Prob-View Excs-Prob-View
Week 2	
R^2	0,46
alpha	1,577
Top 3 3-grams	Pers-Portal-Read Comm-Portal-Read Comm-Portal-Read
	Excs-Score-Memo Excs-Prob-View Excs-Prob-View
	Excs-Prob-View Excs-Score-View Excs-Score-View
Week 3	
R^2	0,672
alpha	1,52
Top 3 3-grams	Exam-Result Exam-Result Exam-Result
	Pers-Portal-Read Comm-Portal-Read Comm-Portal-Read
	Excs-Score-View Excs-Score-View Excs-Score-View
Week 4	
R^2	0,8
alpha	1,559
Top 3 3-grams	Pers-Portal-Read Comm-Portal-Read Comm-Portal-Read
	Exam-Result Exam-Result Exam-Result
	Forum-Read Forum-Read Forum-Read

Furthermore, temporary services and applications have a high impact on the students' learning activities. In our example, the survey which was conducted among students in the holiday season was done by many students via their mobile device. In all seasons except from the post-exam season, solving multiple-choice exercises was the second or third most frequent learning activity sequence. In the post-exam week, by contrast, looking up exam results and forum posts become important sequences. Indeed, exams and their results are often discussed extensively in the LMS forums after the exams. Figure 5 points to a similar direction regarding the volatility of 3-gram learning activities across the four seasons.

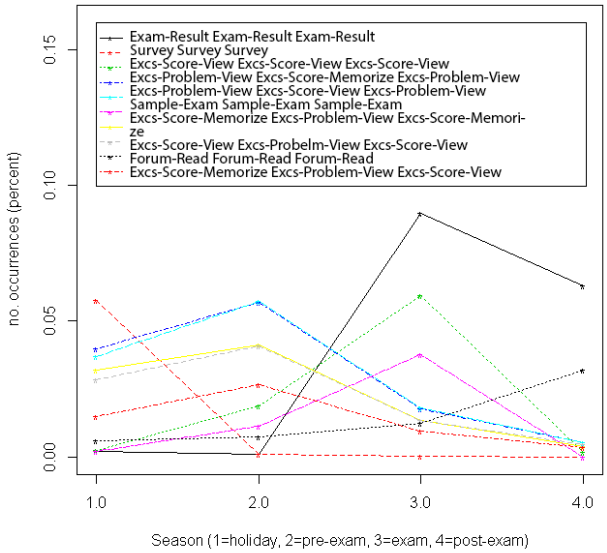

Figure 5: Frequency distribution of 10 most fluctuating learning activities (3-grams) over 4 measurement periods

A drawback of describing activity patterns through *n*-grams is that activity patterns are only recognized if they occur in an exact order within a session. If a user conducts a "deviating" activity during a sequence of activities, his sequence might not be counted as a specific *n*-gram, even if his session is, on a qualitative level, very similar to another session. However, a different sequence might mean a different activity, and since we are interested in the most common occurrences, some random deviations are not of interest of this study. However, aside of sequencing, co-occurrence of activities in the sessions are interesting to identify different types of sessions.

4.4 PCA-based comparison of LMS usage behavior in different seasons

To identify different types of sessions, we applied Principal Component Analysis (PCA) to analyse and compare the session characteristics over the four seasons.

"The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set which consists of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few contain most of the variation present in all of the original variables." [12]

In our case, we started by defining the 1-gram learning activities as the variables. However, the PCA did not automatically lead to a reasonably low number of principal components (PCs), i.e. to a new set of uncorrelated variables, which would explain the characteristics of sessions. Therefore we reduced the number of variables (different types of learning activities) by categorizing these in an inductive process. This resulted in 10 learning activity categories, namely:

- *contribute:* active participation in a course, like posting forum entries
- *inforead:* 'consuming' learning materials
- *infoskip:* briefly looking at learning materials (less than 10 seconds)
- *infosearch:* performing search operations in the platform
- *calendar:* using the calendar module
- *gradeinfo:* looking up grades
- *courseinfo:* looking up information on courses
- *geoinfo:* looking up the location of a course
- *navigate:* making use of the various navigation elements in the LMS platform
- *other:* all other activities

The data sets of the four weeks were adapted and PCA was applied to calculate the PCs for each week.

Table 2: Comparison of the four seasons according to the first five PCs. Cumulative percentages of variance; top-4 activity categories per PC

	Week 1	Week 2	Week 3	Week 4
PC1	24,81%	25,35%	25,91%	24,11%
	contribute (31%)	contribute (32%)	infoskip (24%)	infoskip (31%)
	infoskip (27%)	inforead (28%)	contribute (24%)	inforead (25%)
	inforead (25%)	infoskip (24%)	inforead (24%)	contribute (22%)
	other (9%)	other (12%)	other (16%)	other (11%)
PC2	40,75%	38,16%	38,46%	37,23%
	navigate (44%)	navigate (49%)	courseinfo (37%)	courseinfo (24%)
	courseinfo (24%)	courseinfo (34%)	navigate (27%)	geoinfo (21%)
	gradeinfo (12%)	gradeinfo (9%)	calendar (14%)	calendar (20%)
	other (7%)	calendar (3%)	geoinfo (12%)	infosearch (11%)
PC3	52,83%	49,24%	50,41%	48,76%
	infosearch (42%)	calendar (36%)	gradeinfo (33%)	gradeinfo (24%)
	gradeinfo (22%)	geoinfo (32%)	calendar (32%)	courseinfo (22%)
	other (10%)	gradeinfo (27%)	geoinfo (21%)	calendar (22%)
	infoskip (9%)	navigate (2%)	courseinfo (8%)	geoinfo (8%)
PC4	64,03%	59,48%	60,40%	58,64%
	calendar (82%)	infosearch (62%)	infosearch (77%)	gradeinfo (34%)
	gradeinfo (14%)	gradeinfo (16%)	gradeinfo (13%)	geoinfo (30%)
	courseinfo (2%)	courseinfo (7%)	geoinfo (6%)	other (14%)
	other (1%)	calendar (5%)	contribute (1%)	navigate (9%)
PC5	74,41%	69,04%	69,81%	68,21%
	gradeinfo (35%)	geoinfo (50%)	gradeinfo (33%)	infosearch (40%)
	courseinfo (33%)	gradeinfo (24%)	geoinfo (32%)	navigate (25%)
	infosearch (23%)	calendar (9%)	infosearch (17%)	gradeinfo (20%)
	calendar (3%)	infosearch (5%)	navigate (6%)	calendar (3%)

Table 5 shows the comparison of the four seasons with respect to these PCs. The first principal component (PC1) has a strong focus on the learning activity categories '*contribute*', '*infoskip*', '*inforead*' and '*other*', whereby learning activities indicating an active contribution of users are the most prominent ones in the first two weeks but then start to descend (2nd position in 3rd week and 3rd position in the last week). The second principal component (PC2) consists of activities for navigation, course informations and grades. In this context, users seem to navigate a lot in the LMS and look up course informations in the first two weeks (holidays, pre-exam), while navigation activities descend in the last two weeks and retrieving course informations is getting more important (up to 37%). Finally, it is noticable that the other PCs are dominated by certain activity categories in selected weeks. For instance, PC4 exhibits a high percentage on calendar-related activities (week 1) and information search activities (week 2 and 3).

5. CONCLUSIONS, DISCUSSION AND FUTURE WORK

In this paper we have proposed an educational data mining approach to identify and compare seasonal effects on the basis of patterns of learning activities of users within an LMS platform. By comparing the data-set of several weeks, our case study showed that different seasons can be characterized e.g. through basic statistics on the data-set, rank/frequency plots of sequential activity patterns or types of learning sessions identified via PCA. Consequently, we showed that the intensity of platform usage as well as certain learning activities are highly dependent on the season within an academic year. Although our analysis is limited to four weeks of one semester only, we believe that our approach to detect seasonal effects can be used to improve the platform and for possible personalization and context adaptation features.

With respect to Academic and Learning Analytics, the statistics and activity frequency distribution can be a valuable source for administrators to maintain and improve the ICT infrastructure, i.e. by extending hardware capabilities of the LMS platform and network bandwidth in the busy weeks or by enhancing the usability and performance of the components which are used very intensely. Additionally, seasonal effects can be of interest for the organization, e.g. to identify weeks with less or unused resources, as well as for teachers, e.g. to better distribute the workload over the semester. For these three stakeholder groups of LMS technology a typical Analytics dashboard might be a useful tool for their specific tasks.

With respect to learners, our approach could be used for creating awareness for usage behavior of the current season, e.g. by indicating typical activities and activity. Moreover, season-specific data-sets can be used to generate personalized information [15], such as the most active peer with a similar LMS usage behavior or activities that have been observed in similar browsing sessions. Furthermore, the user interface of the platform can be adapted based on the data, for instance according to typical mechanisms of Adaptive Hypermedia like adaptive navigation support (e.g. the provision of seasonal links) or adaptive presentation techniques [2]. In contrast to other approaches for Collaborative Filtering and Adaptive Hypermedia, personalization would be triggered by characteristics of a season and not by the ones of learners or groups.

Future work should thus focus on further generalizing the classification scheme and on providing this information to learners and other user groups to improve the awareness and the didactical designs. Furthermore, we have investigated four weeks only (i.e. holiday, pre-exam, exam and post-exam). We assume that there are other LMS usage effects that can be used to characterize other kinds of seasons in an academic year. Finally and with respect to mobile computing, we plan to specifically address users who access the LMS with different mobile devices.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231396 (ROLE project). Furthermore, we would like to thank the students of the

"Seminar aus Wirtschaftsinformatik und Neue Medien" at the WU in the summer term 2012, in particular to Martin Bergner, Stephan Feichter and Markus Fill for conducting some of the statistical analyses presented in this paper.

REFERENCES

[1] Bouchet, F., Azevedo, R. and Kinnebrew, J. 2012. Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning. *Proceedings of the 5th International Conference on Educational Data Mining* (Chania, Greece, 2012), 65–72.

[2] Brusilovsky, P. 2001. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*. 11, 1-2 (Mar. 2001), 87–110.

[3] Campbell, J. and Oblinger, D. 2007. *Academic Analytics*. Educause.

[4] Cano, P., Celma, O., Koppenberger, M. and Buldú, J.M. 2006. Topology of music recommendation networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 16, 1 (Jan. 2006), 013107–013107–6.

[5] Ciriani, V., De Capitani di Vimercati, S., Foresti, S. and Samarati, P. 2007. k-Anonymity. *Secure Data Management in Decentralized Systems*. T. Yu and S. Jajodia, eds. Springer-Verlag.

[6] ComScore 2012. *2012 Mobile Future in Focus. Key Insights from 2011 and What They Mean for the Coming Year*. ComScore.

[7] Ferrini, A. and Mohr, J. 2008. Uses, Limitations, and Trends in Web Analytics. *Handbook of Research on Web Log Analysis*. B.J. Jansen, A. Spink, and I. Taksa, eds. IGI Global. 124–142.

[8] Franz, A. and Brants, T. 2006. All Our N-gram are Belong to You. *Research Blog. The latest news from Research at Google*.

[9] Hershkovitz, A. and Nachmias, R. 2011. Online persistence in higher education web-supported courses. *The Internet and Higher Education*. 14, 2 (Mar. 2011), 98–106.

[10] IBM 2004. Log File Formats.

[11] Johnson, L., Adams, S. and Cummins M. 2012. *The NMC Horizon Report: 2012 Higher Education Edition*. The New Media Consortium.

[12] Jolliffe, I.T. 2002. *Principal Component Analysis*. Springer.

[13] Khan, T.M., Clear, F. and Sajadi, S.S. 2012. The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (New York, NY, USA, 2012), 226–229.

[14] Kinnebrew, J. and Biswas, K. 2012. Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. *Proceedings of the 5th International Conference on Educational Data Mining* (Chania, Greece, 2012), 57–64.

[15] Mödritscher, F. 2011. Beyond collaborative filtering: generating local top-n recommendations for personal learning environments. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* (New York, NY, USA, 2011), 41:1–41:4.

[16] Mödritscher, F., Neumann, G. and Brauer, C. 2012. Comparing LMS Usage Behavior of Mobile and Web Users. *2012 IEEE 12th International Conference on Advanced Learning Technologies (ICALT)* (Jul. 2012), 650 –651.

[17] Newman, M. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*. 46, 5 (May. 2005), 323–351.

[18] Schutze, H. and Manning, C.D. 1999. *Foundations of Statistical Natural Language Processing*. Mit Press.

[19] Shawar, B.A. 2009. Learning Management System and its Relationship with Knowledge Management. *Proceedings of the International Conference on Intellignet Computing and Information Systems* (2009), 738–742.

[20] Soper, D.S. and Turel, O. 2012. An *n*-gram analysis of *Communications* 2000–2010. *Commun. ACM*. 55, 5 (May. 2012), 81–87.

[21] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.-N. 2000. Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl.* 1, 2 (Jan. 2000), 12–23.

[22] Stermsek, G., Strembeck, M. and Neumann, G. 2007. A user profile derivation approach based on log-file analysis. *International Conference on Information and Knowledge Engineering (IKE)* (Las Vegas, Jun. 2007).

[23] Venturini, G. 1993. SIA: A Supervised Inductive Algorithm with Genetic Search for Learning Attributes based Concepts. *Proceedings of the European Conference on Machine Learning* (London, UK, UK, 1993), 280–296.

Appendix A

List of 79 learning activities extracted from request URLs:

1. EXCS-PROBLEM-VIEW
2. EXCS-SCORE-VIEW
3. EXCS-PROBLEM-MEMORIZE
4. EXCS-SCORE-MEMORIZE
5. NAVIGATE
6. LOGOUT
7. LOGIN
8. SEARCH
9. SEARCH-NAVIGATE
10. FORUM-POST
11. FORUM-LOOK
12. FORUM-READ
13. FORUM-POST
14. NEWS-LOOK
15. NEWS-READ
16. FAQ-LOOK
17. FAQ-READ
18. SYLLABUS-LOOK
19. SYLLABUS-READ
20. CALENDAR-READ
21. CALENDAR-READ
22. WIKI-LOOK

23. WIKI-READ
24. WIKI-WRITE
25. ADMIN
26. LEARNING-APP
27. PROBLEM-BASED-LEARNING
28. LEARNING-MODULE-LOOK
29. LEARNING-MODULE-READ
30. GLOSSARY-LOOK
31. GLOSSARY-READ
32. BOOK-LOOK
33. BOOK-READ
34. LECTURECAST-LOOK
35. LECTURECAST-READ
36. PERSONAL-PORTAL-CALENDER
37. PERSONAL-PORTAL-CHAT
38. PERSONAL-PORTAL-MAIN
39. PERSONAL-PORTAL-FAQ
40. PERSONAL-PORTAL-FORUMS
41. PERSONAL-PORTAL-FILESTORAGE
42. PERSONAL-PORTAL-NEWS
43. PERSONAL-PORTAL-LECTURECAST
44. PERSONAL-PORTAL-ANNOTATIONS
45. PERSONAL-PORTAL-ASSIGNMENT
46. PERSONAL-PORTAL-GRADEBOOK
47. PERSONAL-PORTAL-HOMEWORK
48. PERSONAL-PORTAL-READ
49. COMMUNITY-PORTAL-READ
50. COMMUNITY-PORTAL-CALENDER
51. COMMUNITY-PORTAL-MAIN
52. COMMUNITY-PORTAL-MEMBERS
53. COMMUNITY-PORTAL-READ
54. COMMUNITY-PORTAL-FAQ
55. COMMUNITY-PORTAL-FORUMS
56. COMMUNITY-PORTAL-READ
57. COMMUNITY-PORTAL-MATERIALS
58. COMMUNITY-PORTAL-NEWS
59. COMMUNITY-PORTAL-READ
60. COMMUNITY-PORTAL-ANNOTATIONS
61. COMMUNITY-PORTAL-LECTURECAST
62. COMMUNITY-PORTAL-LEARNINGMATERIALS
63. COMMUNITY-PORTAL-GRADEBOOK
64. CLUB-PORTAL-READ
65. CLUB-PORTAL-CALENDER
66. CLUB-PORTAL-FORUMS
67. CLUB-PORTAL-MATERIALS
68. CLUB-PORTAL-READ
69. CLUB-PORTAL-NEWS
70. CLUB-PORTAL-MEMBERS
71. CLUB-PORTAL-READ
72. CLUB-PORTAL-FAQ
73. CLUB-PORTAL-READ
74. CLUB-PORTAL-LECTURECAST
75. CLUB-PORTAL-READ
76. COURSE-LISTING
77. COURSE-SEARCH
78. COURSE-LOCATION
79. COURSE-INFO