Interdisciplinary Journal of E-Learning and Learning Objects **Volume 5, 2009**
IJELLO special series of Chais Conference 2009 best papers

Learning about Online Learning Processes and Students' Motivation through Web Usage Mining

Arnon Hershkovitz and Rafi Nachmias
School of Education, Tel Aviv University, Tel Aviv, Israel

ArnonHer@post.tau.ac.il; Nachmias@post.tau.ac.il

Abstract

This study illustrates the potential of applying Web usage mining - the analysis of Web log files - in educational research. It consists of two sub-studies and focuses on two types of analysis, both related to the whole learning process: investigating one learner's activity in order to learn about her or his learning process, and examining the activity of a large group of learners, in order to develop a log-based motivation measure. Subjects were 674 adults who used an online learning unit as part of their preparations for the Psychometric Academic Entrance Exam and whose log files were drawn. The first sub-study aimed to illustrate the knowledge about the online learner that can be extracted from log files, and this resulted in a list of computable, non computable, and higher-level learning variables. In the second sub-study, a log-based motivation measuring tool was developed on the basis of a theoretical framework, a mechanism for computing relevant learning variables, and a clustering of these variables into three groups (associated with the theoretical framework). A discussion of the results, in the context of educational Web mining, is provided.

Keywords: Web-based learning, Web usage mining, log files, online learner, motivation.

Introduction

The Web by now is a firmly established (virtual) reality that offers unprecedented opportunities to education. There are many modes of delivery of online learning (e.g., educational Websites, virtual courses, Web-supported instructional shells, and digital books), providing accessibility to learning materials, facilitating communication among learners and tutors/peers, and possibly helping to improve the learning and teaching process. While using an online learning environment, learners leave continuous hidden traces of their activity in the form of log file records, which document every action taken in three main dimensions: what was the action taken, who took it, and when it was taken. Having this documentation at hand, it is a great challenge to infer from learners' actual behavior as much as we can tell about their learning process. The purpose of this study is to illustrate the potential of the use of Web usage mining as a research methodology in education for extracting information about teaching and learning processes and to gain insights about online learners.

Web usage mining consists of the analysis of logged data of users' activity with the aim of automatically discovering user access patterns. Web usage mining

Editor: Alex Koohang
An earlier, shorter version of this paper was presented at the Chais conference 2009, in Raanana, Israel, and included in Y. Eshet-Alkalai, A. Caspi, S. Eden, N. Geri, & Y. Yair (Eds.), *Proceedings of the Chais conference on instructional technologies research 2009: Learning in the technological era*. Raanana: The Open University of Israel. http://www.openu.ac.il/research_center_eng/conferences.html

has been applied in education research, and a few studies have used it for investigating the behavior of the individual learner and individual differences. This study, which consists of two substudies, aims to illustrate the application of Web usage mining tools and techniques on an educational dataset. It will demonstrate some part of what can be learned about online learners by analyzing their log files, and it does so by referring to two foci.

The first focus is on analyzing one student's activity. Using visualization tools that we developed (*Learnograms*), it is possible to have a thorough and fine-grain look at the activity of one student, in order to investigate his behavior throughout the learning process. This generates some trivial data (e.g., For how long was this student using the system? Which parts of the system did he often visit?), but some more complex observations, too, are possible (e.g., What was the pattern of the student's accessing of the system? How did the student's interest in different parts of the system evolve?). The second focus is on analyzing the activity of a (large) group of students. Learning variables, which were defined using the former methodology, are here calculated for a large population, in order to find similarities between them (i.e., to find groups of variables which behave similarly) or between students (i.e., to find groups of students whose variable values are similar). In this article, we use log files of a fully online course to illustrate how this works on the level of both the single student's behavior and that of a large population. For the latter, motivation was chosen as a high-level variable to be examined.

Background

Educational Web Mining

Web mining – i.e., the application of data mining to data originating from the Web – is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large datasets (Etzioni, 1996; Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The most common kind of Web mining is Web usage mining, the main purpose of which is to discover patterns of usage of Websites by analyzing Web log files, which document every user's access to the site (Cooley, Mobasher, & Srivastava, 1997). Massively used in e-commerce (e.g., by Amazon.com), Web mining – and Web usage mining in particular – is an emerging methodology in education, too (Castro, Vellido, Nebot, & Mugica, 2007; Romero & Ventura, 2007) and has been a focal point of our research group for almost a decade (Nachmias & Hershkovitz, 2006). While the main goal of Web mining (and, in general, data mining) in e-commerce is to increase sales and profit, its goal in e-learning is usually to improve learning/teaching (Zaiane, 2001).

Over the last few years, the use of data mining techniques for researching Web-based and computer-based learning has evolved rapidly, and many studies have been published in the field, including the first book specifically dedicated to this topic: *Data Mining in E-learning* (Romero & Ventura, 2006). Data mining has been applied in different ways, serving as a research tool for answering various educational questions regarding many learning systems (Castro et al., 2007; Romero & Ventura, 2007). In 2008, the first International Conference on Educational Data Mining (EDM'08) was held in Montreal, QC, followed by a second in Cordoba, Spain (July 2009). These conferences were organized by the International Working Group on Educational Data Mining (EDM) (see http://www.educationaldatamining.org). Applying Web mining techniques in education involves four main aspects:

- **Technology**. Both educational systems and data mining applications have technological issues to discuss. Technology-oriented discussions might deal with the essence of data mining application to a specific education system (e.g., Romero, Ventura, & Garcia, 2008) or with the infrastructure needed for logging students' actions more adequately (Romero & Ventura, 2007).

- **Methodology**. Methodology-oriented studies mainly focus on the method of mining rather than on the content logged, and their conclusions will usually refer more to the use of algorithms and less to the potential implications (e.g., Barla & Bielikova, 2007; Freyberger, Heffernan, & Ruiz, 2004; Merceron & Yacef, 2007).
- **Education**. Education-oriented research will focus on the potential benefits of applying a certain technique rather than on the technique itself. Research oriented towards education will heavily use the educational jargon and will discuss methodology mainly in the Methodology chapter (e.g., Baker, Corbett, Koedinger, & Roll, 2006; Cohen & Nachmias, 2006; Sassoon & Nachmias, 1999; Superby, Vandamme, & Meskens, 2006).
- **Ethics/legal**. The ethical and/or legal aspects of using data mining in education research seem to be the most neglected aspect: it is rarely mentioned at all except in a few works dealing solely with this subject (e.g., Rourke, Anderson, Garrison, & Archer, 2001; van Wel & Royakkers, 2004).

Models for applying usage mining as a research methodology in education were suggested by Pahl (2004) and Zaiane (2001), although earlier research already discussed the potential of analyzing online courses using this method (Rafaeli & Ravid, 1997). Regarding the differences between Web mining in education and in e-commerce, Zaiane (2001) stated that the latter aims to transform the surfer into a buyer while the former aims to transform the learner into a more efficient learner. According to Pahl (2004), usage mining in the context of e-learning is totally different from that in e-commerce, since learning is a far more complicated process than shopping, and its cognitive aspects are much more difficult to deduce by means of log files.

In order to describe the variety of applications of Web mining in educational research, we classify them according to two independent parameters: (1) *subject of research* – research might focus on the individual learner or on a group of learners, and (2) *time reference* – a learning process might be analyzed as one uniform unit (i.e., from a summarizing point of view) or at a fine-grain level, describing diverse behaviors/activities that occur in the course of it. The four groups formed by these parameters are:

- *Group view at the end point*. This view may render a bird's eye view of the Website's global usage patterns. The most common variable in Web mining research in education (and in general) under this category is the number of page views which counts the number of times a certain Webpage or the whole Website was entered (e.g., Nachmias & Segev, 2003).
- *Group view of the process*. This view enables understanding of the paths of navigation along the learning process and may shed light on how these paths were formed (e.g., Ravid, Yafe, & Tal, 2002).
- *Individual view at the end-point*. This view may shed light on individual differences in learning-related variables and may be of help in explaining variance among learners (e.g., Talavera & Gaudioso, 2004). The development of the log-based motivation measuring tool presented in this article takes this view.
- *Individual view of the process*. This view, taken by the study presented in this article, offers a qualitative examination of one learner throughout the learning process. Here, the main objective is to understand the learner's behavior during the online learning process by examining qualitative variables such as time patterns manifested while using an educational Website (Hwang & Wang, 2004).

Although online learning has been massively researched, only little was explored regarding the online learner. Web mining techniques provide the researcher with the opportunity of analyzing learners' automatically and continuously documented traces and translating these traces into

meaningful variables that describe the learning process of the online learners; this kind of research comes under the fourth category above and the unprecedented challenge depicts in it is the focal point of this article.

The Online Learner

The use of the Internet as an instructional tool is rapidly increasing world-wide. Literature on online learning addresses, among other things, methods for constructing and managing an online course, ways of improving online teaching, and factors affecting success in online courses. But light is seldom shed on the perspective of the online learner, his or her cognitive characteristics, and the affective aspects of his or her learning process (Picard, et al., 2004).

Research about online learners' activity on the Web usually focuses on operational variables, with attempts to explain individual differences. For example, the variable *time pattern* (measuring the times during which the learner was active) was examined and found to be correlated with achievement (Hwang & Wang, 2004). Another variable is *pace,* found to be correlated with achievement too, as well as being a stable learner's characteristic, independent of content (Clariana, 1990). The *order of contents viewed* was found to be related to thinking processes and learning modes involved in different parts of the online learning environment (Laurillard, 1987). Higher-level variables, describing the characteristics of learners' online learning process, may be found in a small number of studies. These are often divided into two groups: (a) cognitive and metacognitive variables and (b) emotional and motivational variables (American Psychological Association, 1997; Williams, 1993). Attempts have also been made to find correlations between online learning characteristics and learners' affective states (Cohen & Nachmias, 2006; Zaharia, Vassilopoulou, & Poulymenakou, 2004).

It is, of course, not surprising that researching online learners (who often are distant learners) is not an easy task with regards to traditional research methodologies. Web usage mining makes it possible to investigate learners' online behavior, and this is the main challenge of this study (another example for this methodology is presented in Ben-Zadok, Hershkovitz, & Nachmias, 2009). The second sub-study presented here uses motivation to exemplify the investigation of an affective variable, but this sub-study might also be read as a suggested framework for researching other high-level affective or meta-cognitive variables.

Motivation of Online Learners

Motivation has been suggested as a factor explaining individual differences in intensity and direction of behavior (Humphreys & Revelle, 1984). It is generally accepted that motivation is "an internal state or condition that serves to activate or energize behavior and give it direction" (Kleinginna & Kleinginna, 1981). The sources of motivation can be either internal (e.g., interestingness, enjoyment) or external (e.g., wish for high grades, fear of parental sanctions) to the person (Deci & Ryan, 1985). Motivational patterns, in addition to ability, may influence the way people learn: whether they seek or avoid challenges, they persist or withdraw on meeting difficulties, or they use and develop their skills effectively (Dweck, 1986). Different motivational patterns relate to different aspects of the learning process, e.g., achievement goals (performance or mastery), time spent on tasks, performance (Ames & Archer, 1988; Elliott & Dweck, 1988; Masgoret & Gardner, 2003; Singh, Granville, & Dika, 2002).

Unlike configurations in which the instructor sees the students and might infer their motivation level from facial expression, online learning would seem to disable motivation assessment. However, previous research has suggested several methods for tackling this challenge (see de Vicente & Pain, 1998). Table 1 summarizes the motivation-related terms and variables from five studies that mainly used learner-computer interaction data. Following the previously mentioned defini-

tion of motivation and based on the reviewed literature, we suggest a motivation measuring framework which considers three dimensions: (a) engagement - relates to motivation intensity (Although we use the same term, by engagement we mean a more generalized idea than in Beck, 2004 and Cocea & Weibelzahl, 2007); (b) energization, which refers to the way motivation is preserved and directed, and (c) source of motivation (internal or external).

Table 1. Previous research on motivation recognition based on learner-computer interaction

Research	Motivation-Related Terms	Learning Variables from Log Files
Beck (2004)	Engagement (defined by the author)	Question response time; answer correctness
Cocea & Weibelzahl (2007)	Engagement (defined by the authors)	# of pages read; time spent reading pages; # of tests/quizzes; time spent on test/quizzes
Qu & Johnson (2005)	Confidence, confusion, effort	Reading time; decision time (before perform the task); task duration; # of finished tasks; # of tasks performed not from learning "plan"
de Vicente & Pain (2002)	Control, challenge, independence, fantasy; confidence, sensory/cognitive interest, effort, satisfaction.	Quality, speed (of performance), give up
Zhang, Cheng, He, & Huang (2003)	Attention, confidence	# of non-error compilations; ratio of working time and class average; # of hints; # of executions; time until typing in editor

The Study

Within the above-presented framework of educational data mining for researching learners' online behavior and of motivation measuring, the purpose of this study is twofold: (a) we investigate the behavior of the online student while using online learning environment, using data drawn from the system's log files and visualizations of them; and (b) we develop a log-based motivation measuring tool.

In accordance with the two foci presented in the Introduction and so as to implement the two aims of the study, two sub-studies were designed and carried out to illustrate various aspects of Web usage mining application in education research and to present some benefits of this process. The two sub-studies demonstrate a research process consisting of a number of consecutive steps during which data from Web log files are analyzed. The first sub-study, which investigates one student's activity, is a qualitative-type analysis of the logged data. As a result, a set of learning variables is produced, describing the student's behavior and computable from the logged data. Enriching this list of variables might be possible with replicating this process on a larger population (still, a qualitative research). This should yield a set of potentially interesting learning variables which might lead to the examination of individual differences among students. Such an examination is supported by a large-scale analysis of the variables, as will be described in the second sub-study.

For the second sub-study, which focuses on the activity of a (large) group of students', a theoretical framework is first formulated (in our case, resulting in a three-dimensional definition of motivation), a set of learning variables to be associated with this framework is constructed (like in the

first sub-study), and a mechanism for computing these variables in a large population is applied, in order to implicitly calculate them. Once the variables are calculated, a new dataset is formed, in which each student (row) has a tuple of variables (columns) describing her or his behavior. Analysis is then done on this dataset, using Clustering algorithms, so as to investigate the variance among students and discover similarities among variables. This yields groups (clusters) of variables which behave similarly. The basic idea is that variables which behave similarly might be associated with the same dimension of motivation.

Methodology

The Learning Environment

A simple yet very intensive online learning unit was chosen as the research field. This fully-online environment focuses on Hebrew vocabulary and is accessible to students who take a face-to-face preparatory course for the Psychometric Entrance Exam to Israeli universities. The online system is available to the participants from the beginning of the course until the exam date (between 3 weeks and 3 months in total). The system includes a database of about 5,000 words/phrases in Hebrew, where for each word and for each student, a 3-status familiarity of that student with that word is possible: *knows it*, *partially knows it*, or *does not know it*. Status change is carried out by the student and is possible at various stages of the learning process. The system offers the student several alternative learning modes: (a) memorizing, in which the student browses a table of words/phrases along with their meanings; (b) practicing, in which the student browses a table of words/phrases without their meaning; the student may ask for a hint or for the explanation for each word/phrase; (c) gaming; (d) self-testing, in the same format of the exam the students will finally take, and (e) searching for a specific word/phrase.

Population

Log files of 2,162 adults who used the online learning system were analyzed. For the first sub-study, one student was chosen to demonstrate the potential in log file analysis, and his activity was investigated over about two months of using the system (February – April 2007). The second sub-study used data yielded by one month (April 2007). After filtering non-active students and 0-value cases, the research population for sub-study 2 was reduced to N=674.

Log File Description

The researched system logs a student's activity. Thus each student is identified by a serial number (to ensure privacy, the names of the students were removed before starting the analysis). Each row in the log file documents a session, which begins by entering the system and ends with closing the application window. For each session, the following attributes are kept: starting date, starting/ending time, list of actions and their timestamps. Actions documented are every html/asp page in the system, not including actions within Java/Flash applets (i.e., within-game pages are not documented, only the entry to a game and the next action outside it), and this is due to the limitations of the system logging mechanism. Cleaning and preprocessing, the main purpose of which is to prepare the data for initial manipulation and for visualization, were carried out (e.g., removing empty logged records, unifying date format, computing basic variables).

Learnograms

Learnograms are visual representations over time of learning process-related variables. By looking at various *learnograms* different aspects of the learning process can be evaluated, and therefore our main challenge was to develop *learnograms* to cope with different levels of learning

variables. Basic variables are directly derived from the log files (e.g., time, pace, order of contents viewed), and high-level variables should be computed using them and transformed in order to represent both affective and cognitive patterns (e.g., learning strategy, efficiency, anxiety). *Learnogram*-like representations appear in, e.g., Hwang and Wang (2004), and in this article we present a study of a comprehensive analysis of students' behavior using *learnograms* (sub-study 1). The concepts and techniques used in this study serve as the basis for the first phase of the motivation measuring tool which is presented later (sub-study 2). The four basic variables that were chosen to be presented in this study are: (a) *time* – indicates the duration for which the student was logged in to the system (this variable is binary and therefore only the active sessions are shown); (b) *pace* – indicates the pace of using the system in terms of actions (page visits) per minute; (c) *learning modes* – indicates the learning mode in which the student visited, and (d) *perceived knowledge* – indicates the number of words the student marked as known.

Procedure

The two foci of this study – analyzing one student's activity and analyzing a (large) group of students' activity – were separately researched in two sub-studies. For the second sub-study, motivation was chosen as a complex variable that may be calculated on the basis of log files. This section describes the procedure used in the two sub-studies.

Sub-study 1: Identifying a learner's online behavior

One student was chosen for this sub-study, and *learnograms* reflecting his activity were generated. We will call this student *Johnny*. The four *learnograms* of Johnny, representing the four basic variables – time, pace, learning modes, and perceived knowledge – were presented to three education experts in the course of a number of brainstorming meetings. The purpose of these meetings was to list learning variables that might explain variance between students and that are extractable from the log files. Each learning variable was described on three levels: what does it measure, which basic variables relate to it, and how it can be calculated from the related basic ones.

Sub-study 2: Developing a log-based motivation measuring tool

The motivation measuring tool was developed within a framework which consists of four consecutive phases. The first phase includes an explicit and operational definition of the affective features in question; eventually this definition is assessed in view of the empirical results. Next, empirical data are collected, reflecting students' activity in the learning environment examined. These data are analyzed qualitatively (during the second phase), in order to find relevant variables to measure motivation, and then quantitatively (in the third phase) for clustering them according to similarity over a large population. Finally, phase four links the empirical clusters with the theory-based definition. The result of this is a set of variables whose computation is based solely on the log files; at this stage we also relate these variables to the theoretical conceptualization. Below is a detailed description of the phases.

- **Phase I – Constructing a Theory-based Definition**. This phase is based on literature, and it aims to explicitly conceptualize the terms under study. An operational definition regarding motivation is evolved. This definition should later be related to empirical findings.
- **Phase II – Identifying Learning Variables**. This phase is a replication of the methodology presented in sub-study 1. The main purpose of this phase is to find as many learning variables as possible which best reflect motivation. After cleaning and preprocessing the log files, *learnograms* for a few students were observed by education experts in order to find variables indicating individual differences which might be related to the research framework constructed in

the previous phase. A *learnogram* of a basic variable (e.g., visiting different parts of a learning environment) might lead to the need to generate the *learnogram* of a more complex variable (e.g., cumulative activity in a certain part of the system); all of these variables are computed from the log files at a later point in time. A list of variables to be calculated based on the log files is the outcome of this phase.

- **Phase III – Empirically Clustering the Variables**. During this phase, data mining algorithms are applied on a newly formed dataset consisting of the calculated values of the Phase II variables, now for all the students. Then we use a clustering algorithm that groups together different variables by similarity as empirically yielded by the research population (this, though, is not the only possible method and others might be considered).
- **Phase IV – Linking the Empirical Clusters to the Theory-based Definition**. In order to link the empirical clusters with the theory-based definition, data reflecting the students' motivation should be collected (e.g., by questionnaires, interviews, observations, pop-up surveys) and triangulated with the existing log-based data. This can offer a validation of the connection between the empirical data and the definition. In this sub-study, we only present a theory-based validation of the results.

File analysis, *learnograms*, and learning variable computations were all done using Matlab. Clustering analysis was done using SPSS.

Results

Sub-study 1: Identifying the Online Learner Behavior

Four *learnograms* were produced for the following basic variables: time, pace, learning modes, perceived knowledge (presented in Figure 1). These *learnograms* were presented to the experts and served as the basis for the analysis of Johnny's behavior and for the formulation of the learning variables. The learning variables (presented in this section in *italic*) are of four types, reflecting different type of analysis: (a) simple variables, directly extracted from the four basic *learnograms;* (b) computed (both scalars and non-scalars) variables, mainly from the four basic variables (represented in the four *learnograms*); (c) not-computed variables, which are defined here for Johnny, but their computation mechanism for the general case is not yet clear, and (d) higher-level variables which are not well defined (yet). Following is a description of those four types of variables regarding Johnny's activity.

Simple variables

Direct analysis of simple variables can be exemplified by examining the perceived knowledge *learnogram*. We might recall that the perceived knowledge is determined by the number of words/terms the student has marked as known. It is obvious that Johnny's *pace of word marking* is not consistent during his learning period. This variable is quite linear from the beginning until day 33, and then goes through two periods of almost zero value (i.e., no marking at all) – between days 35-48, 49-61 – followed by a high value for some very short periods (zooming-in on the time *learnogram* shows that these high values are a result of one session in both cases). In this manner, a lot can be learned about the learner's behavior from a direct observation of the *learnograms* without any computation.

Computed variables

Following is an example of several computed scalar learning variables. *Total time of being online* is calculated by summing the session durations (yielded by the basic variable of time; a session is a time segment from log-in to log-out; we will not discuss time-out issues in this article).

For Johnny, the value of this variable is 5 hours and 20 minutes. *Number of sessions* is a variable obviously related to the former, and for Johnny its value is 107. Given the session durations, we may obtain Johnny's *average session duration*, which is 3.3 minutes (σ=4.6, longest session was 19.3 minutes). Further examination of Johnny's *learnogram* of time may give us a hint about his *average starting hour of session*. Zooming-in on this *learnogram*, we see that most of his activity is centered on the second half of the day (noon to midnight), and a formal calculation – considering that hour is represented on a [0,24] continuous scale – gives that the average starting hour is 4pm (σ=4.25), i.e., Johnny is an afternoon type of learner.

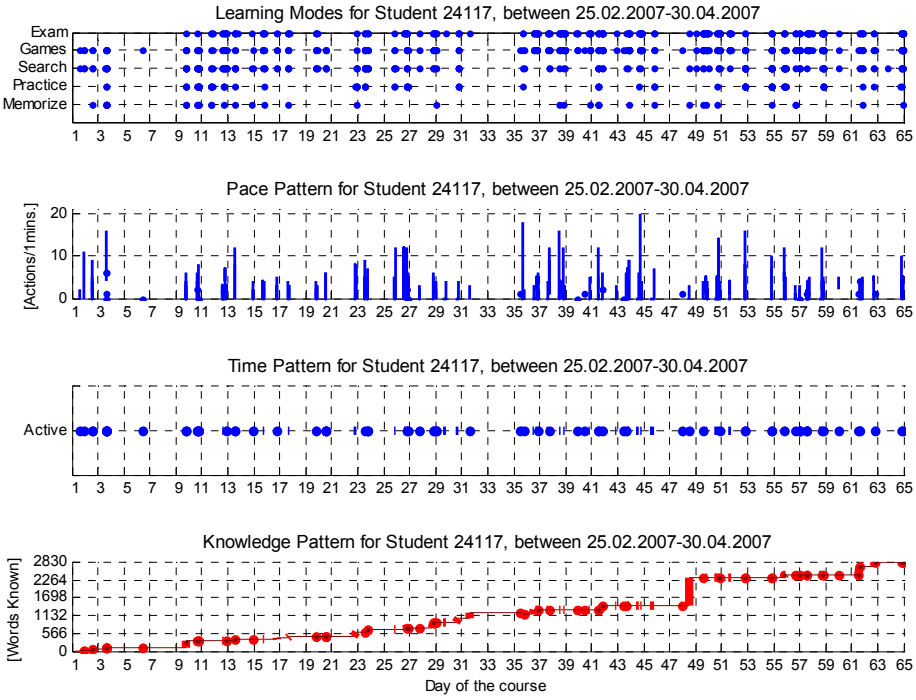

Figure 1. Basic *learnograms* for Johnny (student 24117), for the following basic variables: learning modes, pace, time, perceived knowledge

Looking at the *learnogram* of the basic variable of learning modes, we defined five non-scalar variables, as opposed to the scalar variables, to measure the extent to which each learning mode is being used. They were named *cumulative activity of <learning mode>*, where *learning mode* represents the five learning modes within the system, namely: memorizing, practicing, searching, gaming, taking exams. Each of these variables is a vector of the same length as the four basic variables, consisting of numbers representing the relevant page hits. Therefore, these variables may be visualized using *learnograms* which are not basic but rather computed from basic variables. Figure 2 depicts two of these *learnograms*. We may observe that the pace of the exam activity is quite consistent during the whole learning period, i.e., Johnny uses this mode of learning at the same intensity throughout the course. However, the search activity is not consistent and Johnny uses it mainly between days 1-23, 47-65, while in between there is almost no search activity.

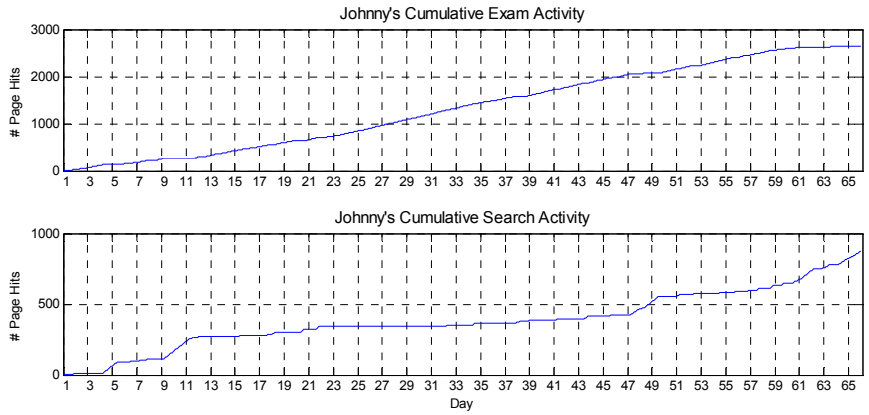

Figure 2. Johnny's *Learnograms* for two computable variables: cumulative exam activity (top), search activity (bottom)

Not-computed variables

Johnny's *learning strategy* illustrates a variable whose calculation mechanism we cannot formally describe (yet). It is based on previously defined and calculated variables. We may see that between days 35-48 Johnny increases his pace of activity in memorizing (days 35-39) and in practicing (days 38-45). Between days 49-61, Johnny simultaneously increases his pace of activity in these two modes (days 52-65). Within those two periods, the pace of gaming and taking exams hardly changes while the search pace slows down dramatically. The search pace increases again towards the end of those two periods and right after them, when Johnny's *pace of word marking* increases steeply. The average *pace* of activity during days 35-65 is higher than the average pace during days 1-35. That is, Johnny's *strategy of learning* changes dramatically during his learning period. First, he chose to mark words as an integral part of the overall activity, but later he chose a totally different strategy of separating the word marking session from other activities. According to this new strategy, he uses the system for 12-13 days during which he focuses on memorizing and practicing and barely marks known words. Subsequently, he devotes an extensive session to word marking during which he makes heavy use of the search engine.

Given the change of strategy, we may suggest that there were three different sub-periods during Johnny's learning period, which may be entitled: *initial contact* (days 1-7, characterized by low activity), *acquaintance and experience* (days 9-32, marking words while using the different modes and by low pace of activity), and *utilization* (days 35-65, a significant change in the learning strategy). This division, based on defined and measured learning variables, renders a very interesting picture of Johnny's behavior (and the changes in it) during the learning period.

Higher-level variables

The real challenge of this study is to find out higher-level educational variables on the basis of previously described variables. For example, the strategy adopted by Johnny for the third subperiod may lead us to an understanding of some higher-level learning variables. Johnny may have an internal *locus of learning control*, a term borrowed based on Rotter's *locus of control* (1966), i.e., Johnny may not need the system to continuously adapt itself according to his word marking, but rather prefers to control it himself. Furthermore, the observed change of strategy during Johnny's learning period may hint that his *motivation* to improve his vocabulary is high, which leads him to improve his way of using the system. This may tell us that Johnny evinces some measure of *learning about his own learning* and that he might have gone through a *reflection process* about his own learning somewhere between days 32-35. These four learning variables are

still not well defined and hence have no computation algorithm. Automating their evaluation process will be possible once we understand their components. Of course, they are yet to be validated regarding the research subject.

Before moving on to the results of the second sub-study, it should be emphasized that the first sub-study was of a qualitative type, while the second one is more of a quantitative type, though both are based on the very same raw data. The second sub-study actually begins where the first one ends, with a list of learning variables in hand. However, for the sake of clarity, we chose to replicate the methodology previously demonstrated in this second sub-study, and it is presented as its second phase, resulting in a different set of learning variables than in the first sub-study, and oriented specifically towards motivation research.

Sub-study 2: Developing a Log-based Motivation Measuring Tool

The four-phase framework described in the Methodology was implemented in the online learning environment investigated, in order to develop a log-based motivation measuring tool. Following is the description of each of the phases.

Phase I – Constructing a theory-based definition of motivation

Based on the reviewed literature, we suggest conceptualizing the motivation measuring tool by reference to three dimensions: (a) engagement - which relates to the intensity of motivation; although we use the same term as Beck (2004) and Cocea and Weibelzahl (2007), what we have in mind is a more general notion; (b) direction - which refers to the way motivation is preserved and oriented, and (c) source of motivation (internal or external). Here it is important to point out that although the variables by which motivation is measured might be (almost) continuously evaluated, motivation – as the sum of many parameters – should be measured over a period of time and is not thought of as a continuous variable. Hence, *engagement* is considered an average intensity, *direction* should describe the overall trend of the engagement level (e.g., increasing, decreasing, stable, frequently changing), and *source* indicates the motivation's tendency to be either internal or external.

Phase II - Identifying learning variables

In order to identify and define motivation-related variables, *learnograms* of the four basic variables – time, pace, learning modes, and perceived knowledge – were produced for five students on the basis of their logged data, reflecting 65 days of activity. As a result of examining these *learnograms*, seven variables were defined and they are detailed in Table 2. These variables are the basis for the analysis in Phase II.

Table 2. The variables defined in Phase II

Variable	Description	Unit
timeOnTaskPC	Total time of active sessions [min] divided by total time of logged data.	[%]
avgSession	Average session duration	[min]
avgActPace	Average pace of activity within sessions; pace of activity per session is the number of actions divided by the session duration	[actions/ min]
avgBtwnSessions	Average time between sessions	[min]

Variable	Description	Unit
wordMarkPace	Pace of word marking: Changed number of known words from beginning to end (can be negative) divided by total time of logged data	[words/ min]
examPC	Percentage of exam-related activity: Number of exam actions divided by total number of actions	[%]
gamePC	Percentage of game-related activity: Number of game actions divided by total number of actions	[%]

Phase III – Clustering the variables empirically

Log files for one month (April 2007) were collected and preprocessed; originally these were the log files of 2,162 students. Students using the researched system were enrolled in different courses (varying in terms of length, intensity, starting date and proximity to the Psychometric entrance exam); however this logged segment was analyzed regardless of students' learning stage. A filter was applied for including students with at least 3 active sessions, leaving about two-thirds of the population (1,444 students). Then, algorithms for calculating the variables were formally written and implemented using Matlab.

First, the variable distributions were examined (see Figure 3). We observed two major problems regarding this distribution which might lead to difficulties in the clustering of variables. The first of these was a significant 0-value noise. This was especially the case for the following three variables: wordMarkPace, examPC, gamePC. Hence, cases with 0-value in either of these variables were cleaned for focusing on the positive-value cases. As a result, the dataset was reduced to its final size, $N = 674$. Second, since we found skewness we used the transformations of log (timeOnTaskPC, avgSession, wordMarkPace, examPC, gamePC) and square-root (avgActPace, avgBtwnSessions).

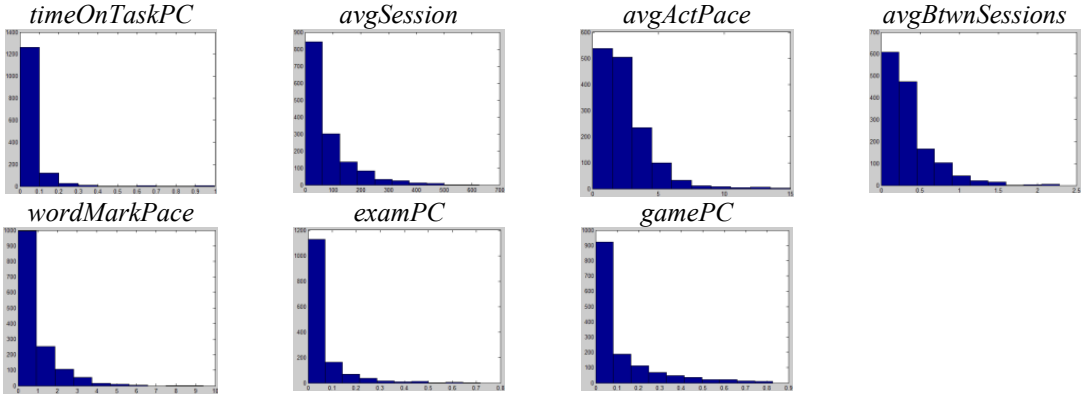

Figure 3. Distribution of the variables, pre cleaning and transformation (*N*=1,444)

Finally, for classifying the variables into groups by similarity, hierarchical clustering of the variables was applied, with Pearson Correlation Distance as the measure and Between Groups Linkage as the clustering method. The clustering process is described by a dendrogram (from the Greek *dendron* "tree", *gramma* "drawing"), presented in Figure 4. The vertical lines describe which variables/clusters were grouped together and at which stage of the algorithm (from left to right). For example, the first coupled variables were timeOnTaskPC and avgSession, and next examPC and gamePC were grouped. According to the results, as shown in the dendrogram, and

for mapping the clusters to our motivation conceptual framework, we decided to define three clusters consisting of the following variables:

- timeOnTaskPC, avgSession
- examPC, gamePC, avgBtwnSessions, avgActPace
- wordMarkPace.

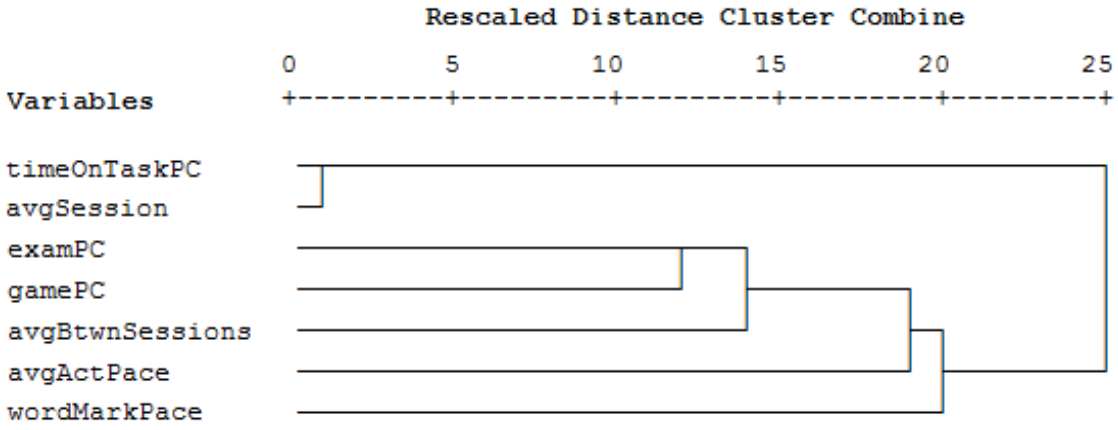

Figure 4. Result of the hierarchical clustering of the variables

Phase IV – Associating the empirical clusters with the theory-based definition

We now suggest a mapping between the empirical clusters and the theory-based definition of motivation, which has served as the conceptual framework for this study. It is important to emphasize that this mapping is currently based on literature review and has not yet been validated. The mapping is described in Table 3.

- Cluster 1. The variables timeOnTask and avgSession, which form the first cluster, might be related to the extent of engagement, as it was previously suggested that working time might be a measure for attention or engagement (Cocea & Weibelzahl, 2007; Dweck, 1986).
- Cluster 2. The variables examPC and gamePC, grouped together in the second cluster, might reflect the students' source of motivation; it may be reasonable to hypothesize – inspired by, e.g., Heyman & Dweck (1992) and Ryan & Deci (2000) – that students who tend to take self exams frequently (related to performance goal orientation) have extrinsic motivation to learn, while those who tend to game applications (related to learning goal orientation) are intrinsically motivated. The variables avgActPace and avgBtwnSessions are also clustered together with the previous two, but their closeness to source of motivation is yet to be established.
- Cluster 3. The variable wordMarkPace, indicating students' word-marking speed, forms the third cluster. According to a diagnostic rule found in de Vicente and Pain (2002), high speed of activity together with high quality of performance (when staying in similarly-difficult exercises) suggests increasing motivation. Since an increase in the number of words marked is, to some extent, an indication of the student's knowledge (i.e., a reflection of performance), wordMarkPace might be related to the direction of motivation, i.e., direction.

Table 3. The resulting clusters and their mapping to the motivation dimensions

Cluster	1	2	3
Variables	*timeOnTaskPC* *avgSession*	*examPC* *gamePC* *avgBtwnSessions* *avgActPace*	*wordMarkPace*
Motivation dimension	Engagement	Source	Energization

It is clear that validating these results and scaling the variables are crucial before completion of the development of the motivation measuring tool. The proper way of doing this is by an external validation, i.e., identifying the association between the variables found and independent variables measured by an external measuring tool for motivation. It is also possible to examine the validation step by referring to a different learning environment; however, in this case a few preliminary steps are required, particularly a replication of the clustering process, in order to ensure that the new system preserves the found clusters.

Discussion

Although much educational research was done with Web mining methodologies, only a few studies may be categorized as analyzing the individual learner's behavior during the whole learning process. This is not surprising, as researching a student in a non face-to-face learning scenario is not an easy task if we use traditional research methodologies. However, while learning online, students leave continuous and very detailed traces of their activity. Using these traces – kept in the form of log files – for investigating the learners' behavior has a great potential for serving different aspects of educational research (Castro et al., 2007; Romero & Ventura, 2007).

The challenge of applying Web usage mining – i.e., the analysis of Web log files for discovering patterns within them – is twofold. On one hand, we aim to learn as much as possible about the individual learner, and on the other, we aim to generalize this knowledge to a large population. For tackling the first challenge, we developed the *learnogram*, a visual representation of learning process-related variables over time. *Learnograms* may present basic variables directly derived from the log files, as well as higher-level variables based on previously defined variables. Substudy 1 offered a detailed presentation of the analysis of the learnograms of one learner. The investigation was based on the choice of four basic variables, extracted directly from the log file. Since the basic variables are the basis for forming the other variables, they should be examined and might be changed. However, we feel that the basic variables defined here (excluding perceived knowledge) are quite straightforward, essential for any analysis, and extractable from almost any log file.

The second challenge was demonstrated in sub-study 2, during which a motivation measuring tool was developed, based solely on log files. It is necessary to validate the results of this process using external measures of motivation; however, two additional major limitations are to be considered. First, variables were identified in a specific learning environment; the measuring tool, hence, might be useful for similar systems, but when using it in different environments (in terms of, e.g., learning domain, instruction modes available) they should be converted and their clustering should be re-examined. Second, the tool might be incomplete; we only focused on seven variables but others might be considered. Identifying these variables from a segment of the learning – as was demonstrated here, since the log files did not necessarily reflect a whole learning process

from beginning to end – makes it possible to employ this tool during the learning process; in this way, intervention, when needed, may be possible and changes in motivation may be analyzed.

Besides demonstrating the idea that qualitative-type – and not only quantitative-based – knowledge about online learners might be extracted from their log files traces, as well as high-level variables describing the learning process, this study also showed the other side of the coin: this is a very difficult task. The process of using Web usage mining in itself is complex; however, when applying this methodology in the context of education research, the task becomes even more difficult, as the gap formed between the students' action – represented in the log files – and their cognitive, meta-cognitive, and affective behavior during the learning process requires the building of bridges. Nevertheless, the potential of this method might be huge, as it may reveal new perspectives on many aspects of the learning/teaching.

Any discussion concerning the application of Web mining in educational research should not ignore the ethical, and sometimes even legal, aspects of this methodology. Although it seems that once the raw data were fully randomized, the students who originated these logs were free of any harm, questions regarding ethical issues – such as formal consent, privacy, and de-individualization (Rourke et al., 2001; van Wel & Royakkers, 2004) – should be thoroughly discussed.

Web mining research in education and, in general, educational data mining is a multidisciplinary field bringing together scholars from many disciplines, mainly from education, computer sciences, information sciences, psychometrics, psychology, and statistics. Since every discipline has its own jargon, communication between scholars is not trivial. We feel that evolving a common language within the community is a great challenge that might help in promoting the research in the field. This is one of the purposes of the International Working Group on Educational Data Mining (EDM) and also an important objective of our EduMining research group (http://edumining.info) within the Knowledge Technology Lab. Regarding this objective, as well as where it comes to gaining knowledge about the online learner by using Web mining techniques, it seems that these are only our first steps in a challenging and promising thousand-mile journey.

References

American Psychological Association. (1997). *Learner-centered psychological principles: A framework for school reform*. Washington, DC.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation. *Journal of Educational Psychology, 80*(3), 260-267.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., & Roll, I. (2006). Generalizing detection of gaming the system across a tutoring curriculum. In M. Ikeda, K. Ashlay, & T.-W., Chan (Eds.), *Intelligent Tutoring Systems: Proceedings of the 8th International Conference, ITS 2006*, Jhongli, Taiwan, June 26-30, 2006.

Barla, M., & Bielikova, M. (2007). Estimation of user characteristics using rule-based analysis of user logs. *Proceedings of the Workshop on Data Mining for User Modeling, 11th International Conference on User Modeling, Corfu, Greece*.

Beck, J. E. (2004). Using response times to model student disengagement. In *Proceedings of the Workshop on Social and Emotional Intelligence in Learning Environments, 7th International Conference on Intelligent Tutoring Systems, Maceio, Brazil*.

Ben-Zadok, G., Hershkovitz, A., & Nachmias, R. (2009). A case study of assessing learning process in Web-based science learning environment using visual representation of log files. In A. Méndez-Vilas, A. Solano Martín, J.A. Mesa González & J. Mesa González (Eds.), *Research, reflections and innovations in integrating ICT in education*. Badajos, Spain: FORMATEX.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In L. C. Jain, T. Raymond & D. Tedman (Eds.), *Evolution of teaching and learning paradigms in intelligent environment* (Vol. 62, pp. 183-221). Berlin: Springer-Verlag.

Clariana, R. B. (1990). Rate of activity completion by achievement, sex and report in computer-based instruction. *Journal of Computing in Childhood Education, 1*(3), 81-90.

Cocea, M., & Weibelzahl, S. (2007). Cross-system validation of engagement prediction from log files. In E. Duval, R. Klamma, & M. Wolpers (Eds.), *Creating New Learning Experiences on a Global Scale: Proceedings of the Second European Conference on Technology Enhanced Learning, EC-TEL 2007*, Crete, Greece, September 17-20, 2007.

Cohen, A., & Nachmias, R. (2006). A quantitative cost effectiveness model for Web-supported academic instruction. *The Internet and Higher Education, 9*(2), 81-90.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *Proceedings of the IEEE 9th International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, CA*.

de Vicente, A., & Pain, H. (1998). Motivation diagnosis in intelligent tutoring systems. In B. P. Goettl, H. M. Halff, C. Redfield, & V. Shute (Eds.), *Proceedings of Intelligent Tutoring Systems: 4th International Conference, ITS'98, San Antonio, Texas, USA*, August 16-19, 1998.

de Vicente, A., & Pain, H. (2002). Informing the detection of the students' motivational state: An empirical study. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of Intelligent Tutoring Systems: 6th International Conference, ITS 2002, Biarritz, France and San Sebastian, Spain*, June 2-7, 2002.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040-1048.

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*(1), 5-12.

Etzioni, O. (1996). The World Wide Web: quagmire or gold mine? *Communications of ACM, 39*(11), 65-68.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Freyberger, J., Heffernan, N. T., & Ruiz, C. (2004). Using association rules to guide a search for best fitting transfer models of student learning. In *Proceedings of the Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes, 7th International Conference on Intelligent Tutoring Systems, Maceio, Brazil.*

Heyman, G. D., & Dweck, C. S. (1992). Achievement goals and intrinsic motivation: Their relation and their role in adaptive motivation. *Motivation and Emotion, 16*(3), 231-247.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91*(2), 153-184.

Hwang, W.-Y., & Wang, C.-Y. (2004). A study of learning time patterns in asynchronous learning environments. *Journal of Computer Assisted Learning, 20*(4), 292-304.

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5*(4), 345-378.

Laurillard, D. (1987). Computers and the emancipation of students: giving control to the learner. *Instructional Science, 16*(1), 3-18.

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning, 23*(1), 123-163.

Merceron, A., & Yacef, K. (2007). Drawbacks and solutions of applying association rule mining in learning management systems. In *Proceedings of the Workshop on Applying Data Mining in e-Learning, the 2nd European Conference on Technology Enhanced Learning (EC-TEL'07), Crete, Greece.*

Nachmias, R., & Hershkovitz, A. (2006). Learning about the online learner. In *Proceedings of the Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection, the 15th International World Wide Web Conference, Edinburgh, Scotland*.

Nachmias, R., & Segev, L. (2003). Students' use of content in Web-supported academic courses. *The Internet and Higher Education, 6*(2), 145-157.

Pahl, C. (2004). Data mining technology for the evaluation of learning content interaction. *International Journal of E-Learning, 3*(4), 47-55.

Picard, R., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., et al. (2004). Affective learning — A manifesto. *BT Technology Journal, 22*(4), 253-269.

Qu, L., & Johnson, W. L. (2005). Detecting the learner's motivational states in an interactive learning environment. In C.-K. Looi, G. I. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Artificial Intelligence in Education - Supporting Learning through Intelligent and Socially Informed Technology, Proceedings of the 12th International Conference*.

Rafaeli, S., & Ravid, G. (1997). Online, web based learning environment for an information systems course: Access logs, linearity and performance. In *Proceedings of the Information Systems Education conference (ISECON '97), Orlando, Florida, USA*.

Ravid, G., Yafe, E., & Tal, E. (2002). Log files as an indicator of online learning and as a tool for improving online teaching. In *Proceedings of the Internet Research 3.0, Maastricht, The Netherlands*.

Romero, C., & Ventura, S. (2006). *Data mining in e-learning*. Southampton, UK: WIT Press.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*, 1-28.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. International *Journal of Artificial Intelligence in Education, 12*, 8-22.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54-67.

Sassoon, E., & Nachmias, R. (1999). What makes them click: How learners utilize Web based courses. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2009), Seattle, WA*.

Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *Journal of Educational Research, 95*(6), 323-332.

Superby, J. F., Vandamme, J.-P., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In M. Ikeda, K. Ashlay, and T.-W., Chan (Eds.), *Proceedings of the Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan*, June 26-30, 2006.

Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *Proceedings of the Workshop on Artificial Intelligence in Computer Supported Collaborative Learning, European Conference on Artificial Intelligence, Valencia, Spain*.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology, 6*(2), 129-140.

Williams, M. D. (1993). *A comprehensive review of learner-control: The role of learner characteristics*. Paper presented at the Annual convention of the Association for Educational Communications and Technology, New Orleans, LA.

Zaharia, P., Vassilopoulou, K., & Poulymenakou, A. (2004). Designing affective-oriented e-learning courses: An empirical study exploring quantitative relations between usability attributes and motivation to learn. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2004), Lugano, Switzerland*.

Zaiane, O. R. (2001). Web usage mining for a better Web-based learning environment. In *Proceedings of the 4th IASTED International Conference on Advanced Technology for Education (CATE'01), Banff, Canada*.

Zhang, G., Cheng, Z., He, A., & Huang, T. (2003). A WWW-based learner's learning motivation detecting system. *Proceedings of the International Workshop on Research Directions and Challenge Problems in Advanced Information Systems Engineering, First International Conference on Knowledge Economy and Development of Science and Technology, Honjo City, Japan*.

Biographies

Prof. Rafi Nachmias is a member of the Science Education Department in Tel-Aviv University's School of Education. He is currently heading the Science and Technology Education Center and the Virtual TAU project in Tel-Aviv University, and also serves as Vice Head of School of Education for research. His major research areas are: Internet and Higher Education, Web-Mining of online learning, Web-based Learning and Innovative pedagogical school practices using ICT.

Arnon Hershkovitz is a PhD student in the School of Education at Tel Aviv University. His research aims at applying data mining techniques for enriching the knowledge about the behavior of online learners. He holds a BA in Mathematics and Computer Science and an MA in Applied Mathematics. Arnon is an active member in the International Working Group on Educational Data Mining.