A Case Study of Using Visualization for Understanding the Behavior of the Online Learner

Rafi Nachimas, Arnon Hershkovitz

Science and Technology Education Center, School of Education, Tel Aviv University,
Tel Aviv 69978, Israel
{nachmias, arnonher}@post.tau.ac.il

Abstract. This paper describes a case study of the behavior of an online learner by visualizing log file records using a tool (learnogram) we developed. In this study, we chose a simple yet very intensive fully-online learning environment (for learning Hebrew vocabulary) and one student who uses it. With the help of brainstorming meetings with three education experts, fifteen learning variables were listed; some of them were formally defined and calculated. We conclude the paper with a discussion about the challenges this method raises and its great potential towards the development of adaptive Web-based learning environments.

1 Introduction

The Web nowadays is a firmly established (virtual) reality that offers unprecedented opportunities to education. Many modes of delivery of online learning exist (e.g. educational software, virtual courses, blended learning, electronic books), all 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 by three parameters: what was the action taken, who took it and when. The main objective of the case study presented in this paper is to extract learning-related variables from raw log files using Web mining techniques and a special visualization tool, learnogram.

Web mining is a field consisting of data mining techniques that automatically discover and extract information from Web files. Massively used in e-commerce (e.g., in Amazon.com), Web mining is an emerging methodology also in education [1], and is a focal point of our research group for almost a decade [2].

The process of translating raw log files into meaningful information about the behavior of the online learner - a field not deeply explored yet - is significant, first and foremost, for understanding the essence of online learning. Having the ability to automatically identify learning-related information while the learning process occurs is meaningful for instructors, developers and policymakers. Therefore, the focus of this case study, which is part of a larger research which deals with applying Web

mining techniques in education, is on examining the method of extracting learning related variables by visualizing the raw log files data.

2 Web Mining in Education

The term Web mining (Web data mining), was first mentioned by Etzioni [3], who suggested that traditional data mining techniques for finding hidden patterns in huge databases, can be applied to Web-based information. Web mining is an emerging methodology in education research, assisting instructors and developers in improving learning environments and supporting decision-making of policymakers [4].

Models for applying usage mining as a research methodology in Education were suggested by Pahl [5] and Zaiane [6], although earlier research already discussed the potential of analyzing online courses using this method [7]. Regarding the differences between Web mining in Education and in e-commerce, Zaiane stated that the forlatter aims on transforming the surfer into a buyer while the former aims on transforming the learner into a more efficient learner. According to Pahl, usage mining of e-learning is totally different from usage mining of e-commerce, since the learning process is far more complicated than the shopping process, and its cognitive aspects are much more difficult to track by means of log files.

In order to describe the variety of applications of Web mining in educational research, we classify them into four categories according to the number of learners involved in the research (one learner or a group of learners) and the point of view the research takes (examining the learning process at its ending-point or throughout it). A detailed description of those categories is given in [8], and here we describe them briefly: a) *Group view at the ending point of the learning process* 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., [9]); b) *Group view of the whole learning process* enables to understand the paths of navigation along the learning process, and may shed light on how these paths were formed (e.g., [10])*;* c) *Individual view at the ending-point of the learning process* may shed light on individual differences in learning-related variables, and may be of help in explaining variance among learners (e.g., [11]); d) *Individual view of the whole process* – the angle that the case study presented in this paper takes – is the mode that offers a qualitative picture of one learner throughout the learning process. Here, the main objective is an understanding of the learner behavior during the online learning process, by examining qualitative variables, such as time patterns manifested in an educational Website [12]. This way, analyzing the log files enables us to virtually view this learner, as if we were watching him from aside. The tool we develop – learnogram – promotes the understanding of the online learner's behavior, by visualizing learning variables (see section *Methodology*).

Although online learning has been massively researched, only little was explored regarding the online learner. This is of no surprise: traditional research methodologies can hardly cope with gathering of information about the distant online learner. Web mining techniques provide the researcher with the opportunity of collecting the

learners' traces, which are documented automatically and continuously. Web mining algorithms might enable the researcher to translate these traces into meaningful variables that describe the learning process of the online learners. This is an unprecedented challenge, and therefore it is the focal point of this research.

3 The Online Learner

The use of the Internet as an instructional tool is rapidly increasing, with millions of learners in the United States [13] and meaningful presence in Israel too [14, 15]. A variety of online learning modes is available, such as: fully-online courses, where most of content (usually more than 80%) is delivered online, and typically have no face-to-face meetings [16]; virtual learning communities, in which learners discuss relevant issues with peers and/or instructors and may conduct meaningful collaborative activities [17]; or blended learning, in which a combination of face-to-face instruction and online attendance is offered [18].

The 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 seldom light is shed on the perspective of the online learner, his or her cognitive characteristics and the affective aspects of his or her learning process [19].

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" (trying to measure the times during which the learner was active) was examined and found to be correlated with achievement [12]. Another variable is pace, which was found to be correlated with achievement, as well as being a stable learner's characteristic, independent of content [20]. 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 [21].

Higher-level variables, describing the characteristics of learners' online learning process, may be found in a few studies. These are often divided into two groups: a) cognitive and metacognitive variables; and b) emotional and motivational variables [22, 23]. Attempts have also been made to find correlations between online learning characteristics and affective states of the learner [4, 24].

The objective of the case study presented in this article is to examine the method of extracting different kinds of learning related variables from the raw data documented in the log files, using learnograms – a visualization tool we developed.

4 Methodology

This article presents a case study of analyzing one student's log files from a specific learning environment (which will be presented in details in the next section). The main objective of this case study is to understand what kind of learning variables might be extracted from the raw log files, using learnograms. It is a part of a larger

research, aiming on exploring the essence of the online learning process of the online learner, using information stored in log files and Web mining techniques.

4.1 Research Field

A simple yet very intensive online learning unit was chosen as the research field. This fully-online environment, which focuses solely on Hebrew vocabulary, is accessible for students who take a (face-to-face) course preparing them for the Psychometric Entrance Exam. The material being taught in that online unit is not being taught in class and students who choose not to take the online unit acquaint it with a book.

Log files of this environment document a large part of the activities available in the system (including client-side logging), therefore offering a broad view of the learners' activity. Each year, about 10,000 students (between the ages 18-25) from all over Israel enroll in these courses, and will potentially use the software.

The system holds a database of around 5,000 words/phrases in Hebrew the student should learn. The modes of learning are varied: a) *memorizing* – the student browses a table of the words/phrases with their meaning, and tries to memorize it; b) *practicing* – the student browses the table to the words/phrases, and checks whether he or she knows their meaning. The student may ask for a hint or for the solution; c) *searching* – the student can search for specific words/phrases from the database; d) *gaming* – the student plays games which aim on teaching him or her the words/phrases in an experiential way; e) *exam* – the student takes self-exams which are built according to the real exam they would finally take.

While using the different modes of learning, the student may mark each word/phrase as "well known", "not-well known" or "unknown". During the memorizing and practicing modes, the system presents to the student only those words which he or she didn't mark as "known".

4.2 Learnograms

The main tool that promotes our understanding of the online learner’s continuous behavior is the learnogram. It is inspired by the electrocardiograms (ECG), which charts heart activity. Just as the cardiologist examines ECG charts and is able to describe the patient's heart condition, we aim to understand the learning processes in which the online learner is involved, only by looking at his or her learnograms.

Learnograms are visual representations of learning process-related variables over time. Looking at various learnograms, different aspects of the learning process will be evaluated, and therefore our main challenge is to develop learnograms to cope with difference levels of learning variables. Basic variables are directly derived from the log files (e.g. time, pace, order of contents viewed), 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).

In this case study, four basic variables were chosen, and their learnograms were generated: a) *time* – indicates the time during 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 by terms of actions (page visits) per minute; c) *learning modes* – indicates the learning mode (see *4.1 Research Field*) in which the student visited; d) *knowledge* – indicates the number of words the student marked as known (see previous section).

4.3 Procedure

Log files from the learning environment were collected for the period of February-April 2007. Among the students documented in these files, one student was randomly chosen, and learnograms reflecting his activity were generated. We will call that student *Johnny*.

The learnograms of this student were presented to education experts (N=3) and brainstorming meetings with them were hold. Each learning variable was described in three levels: a) what does it measure; b) which basic variables relate to it; and c) how can it be calculated from the related basic ones. File analysis, learnogram drawings and learning variable computations were all done using Matlab.

5 Results

At the first stage, four learnograms were produced for the basic variables: time, pace, learning modes, knowledge (presented in Figure 1). Those learnograms were presented to the experts and served as the basis for the analysis of Johnny's behavior and for the formation of the learning variables. The learning variables (presented in this section in *italic*) are based on four types of analysis: a) direct analysis extracted from the four basic learnograms; b) computed (both scalars and non-scalars) learning variables which are calculated from the basic variables; c) non computable learning variables which are defined for Johnny, but their computation mechanism is not yet clear for the general case; d) higher-level variables, which are not well defined (yet). Following is a description of those four types of analysis regarding Johnny's activity.

Direct Analysis
An example to a direct analysis is given by examining the knowledge learnogram. It is obvious that Johnny's *pace of words marking* is not consistent during his learning period. This variable is quite linear from the beginning and until day 33, and then has 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 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 from a direct observation of the learnograms about the learner's behavior without any computation.

Computed Learning Variables
Following is an example of several scalar computed learning variables. *Total time of being on-line* is calculated by summating the overall session durations (given by the basic variable time), and for Johnny its value is 5 hours and 20 minutes. (A session is

a time segment from log-in to log-out; we will not discuss time-out issues in this article). *Number of sessions* is an obvious variable related to the former, and for Johnny its value is 107. Having 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 time basic learnogram may hint us about his *average hour of session starting*. Zooming-in this learnogram, it is clear that most of his activity is centered on the second half of the day (noon to midnight), and a formal calculation gives that the average starting hour is 4pm (σ=4.25), i.e. Johnny is an afternoon type of learner.

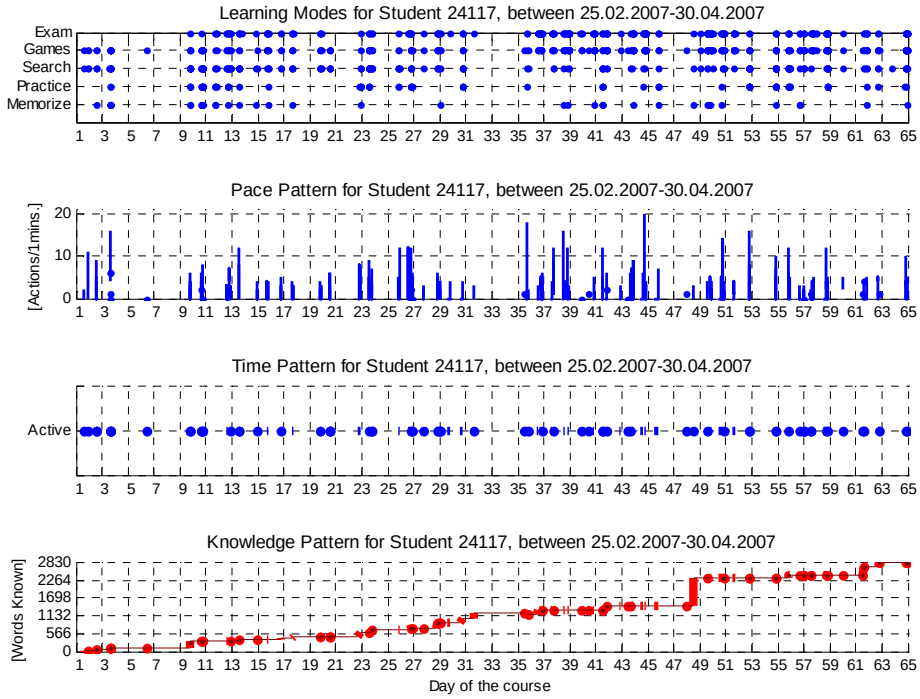

Figure 1 - Johnny's basic learnograms: learning modes, pace, time, knowledge

As opposed to the scalar variables, looking at the learnogram of the basic variable learning modes we have defined five non-scalar variables to measure the extension to which each learning mode is being used. They were named *cumulative activity of [memorizing, practicing, searching, gaming, taking exams]*. Each of those variables is a vector of the same length of the four basic variables consisting of numbers representing the relevant page hits. Therefore, these variables may be visualized using learnograms which are not a basic, but rather computed from basic variables. In Figure 2, two of those learnograms are shown. 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 in the same intensity all over the course. However, the

searching activity is not consistent and Johnny uses it mainly between days 1-23, 47-65, while in between there is almost no searching activity.

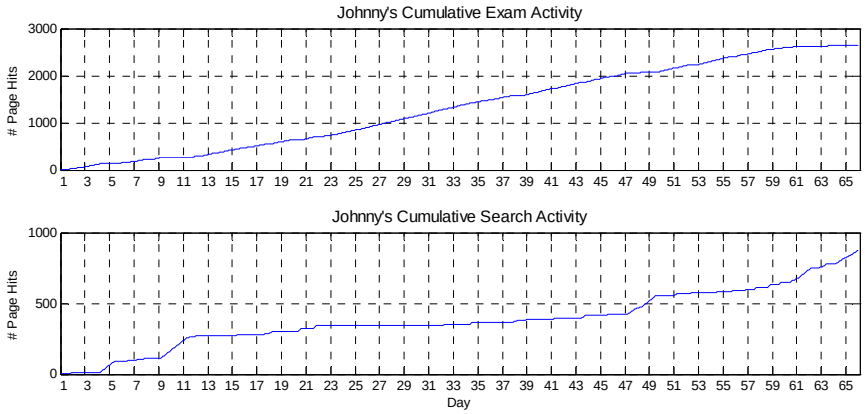

Figure 2 – cumulative exam (top) and search (bottom) activity of Johnny

Non Computable Learning Variables

Johnny's *strategy of learning* is an example to a variable which we cannot formally describe its calculation mechanism (yet). It is based on already 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 almost doesn't change while the searching pace is dramatically slowing down. The searching pace is increasing again towards the end of those two periods and right after them, when Johnny's *pace of words marking* is dramatically increasing. The average *pace* of activity during days 35-65 is higher than the average pace during the days 1-35. That is, Johnny's *strategy of learning* has been dramatically changed 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 words marking session from the 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. Afterwards, he devotes an extensive session to words marking during which he heavily uses the search engine.

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

Higher-Level Variables
The real challenge of this work is to find out higher-level educational variables, based on the previously described variables. For example, the strategy adopted by Johnny for the third sub-period may lead us to the understanding of some higher-level learning variables. It is possible that Johnny has an internal *locus of learning control* (a term that should remind the locus of control [25]), i.e. he doesn't need the system to continuously adapt itself according to his own words marking, but rather prefers his own control on it. Furthermore, the change of strategy during Johnny's learning period may hint that his *motivation* to improve his vocabulary is high, and therefore he improves his way of using the system. This may tell us that Johnny has 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 hence have no computation algorithm. Automating their evaluation process will be possible upon understanding their components.

6 Discussion

Web mining - a field consisting of data mining techniques for discovering and extracting information from Web files - is an emerging methodology also in education. Although many researches have been done in this area, only few may be categorized as analyzing the individual learner's behavior during the whole learning process (for the full categorization, see [8]). For doing this, 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 already defined variables.

The case study presented here demonstrates the method of using the learnograms for understanding the behavior of an individual learner over time. This case study of only one student using one particular Web-based learning environment demonstrates the challenges in our current larger research:

1. *Define and compute as many learning variables as possible*. We focus on the most important variables reflecting the online learners' behavior from the educational point of view. For doing this, we will conduct some further case studies. Variables should be well defined and present with a clear computation mechanism. Since the basic variables are the basis for the other variables, they should be examined and might be changed. However, we feel that the basic variables defined here (excluding *knowledge*) are quite straightforward and essential for any analysis.
2. *Describe the learning variables distribution over large populations.* The learning variables may help us with identifying individual differences between online learners, hence stepping forward to a better understanding of the essence of the online learning of different learners. For implementing this and the former challenge, we do need some better visualization tools.
3. *Extract high-level learning patterns*. Having in hand a list of well-defined and computable learning variables, we may extract higher-level (e.g., meta-cognitive, affective) learning patterns. Our way of doing it will be by using advanced statistical methods (e.g., Cluster Analysis, Decision Trees) in order to identify

interesting patterns in those variable expressions over large learner populations (i.e., the patterns will be of *variables* and not of learners).

4. *Evaluate the transferability of this methodology*. For many applications of this research, we will have to make sure the whole process – from the learnograms presentation and till the high-level variables clustering – is transferable to any Web-based learning environment. This might require a formal system-independent description of this methodology to be evaluated by other researchers.
5. *Outline ethical and legal principles*. The online learner may be unaware that private information is continuously being traced and recorded, stored and analyzed using implicit methods. We are intending to shed light on those concerns, in order to present with some appropriate solutions (for researchers, as well as for learners, online learning system developers and policymakers).

As reported in this paper, we are at the beginning of a long way. The cardiologist examining the patient's ECG may easily identify cardiac behavior. We are still very far from bringing this ordinary procedure to the online learning realm. We do believe that coping with this challenge will enable instructors to identify learning behavior of their students. Moreover, this kind of identification may be supported by the system, hence stepping forward towards adaptive Web based learning environments.

References

1. C. Romero and S. Ventura, Data mining in e-learning. Southampton, UK: WIT Press, 2006.
2. R. Nachmias and A. Hershkovitz, "Learning about the online learner," presented at Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection (in WWW'2006), Edinburgh, Scotland, 2006.
3. O. Etzioni, "The World Wide Web: quagmire or gold mine?," Communications of ACM, vol. 39, pp. 65-68, 1996.
4. A. Cohen and R. Nachmias, "A quantitative cost effectiveness model for Web-supported academic instruction " The Internet and Higher Education vol. 9, pp. 81-90, 2006.
5. C. Pahl, "Data mining technology for the evaluation of learning content interaction," International journal of E-Learning, vol. 3, pp. 47-55, 2004.
6. O. R. Zaiane, "Web usage mining for a better Web-based learning environment," presented at 4th IASTED International Conference on Advanced Technology for Education (CATE'01), Banff, Canada, 2001.
7. S. Rafaeli and G. Ravid, "Online, Web based learning environment for an Information systems course: Access logs, linearity and performance," presented at Information Systems Education Conference, Orlando, FL, 1997.
8. R. Nachmias and A. Hershkovitz, "Web usage mining in online learning: From global to local view," Unpublished manuscript, 2007.
9. R. Nachmias and L. Segev, "Students' use of content in Web-supported academic courses," The Internet and Higher Education, vol. 6, pp. 145-157, 2003.
10. G. Ravid, E. Yafe, and E. Tal, "Log files as an indicator of online learning and as a tool for improving online teaching," presented at Internet Research 3.0, Maastricht, The Netherlands, 2002.

11. L. Talavera and E. Gaudioso, "Mining student data to characterize similar behavior groups in unstructured collaboration spaces," presented at Workshop on Artificial Intelligence in Computer Supported Collaborative Learning at European Conference on Artificial Intelligence, Valencia, Spain, 2004.
12. W.-Y. Hwang and C.-Y. Wang, "A study of learning time patterns in asynchronous learning environments," Journal of Computer Assisted Learning, vol. 20, pp. 292-304, 2004.
13. I. E. Allen and J. Seaman, "Growing by Degrees: Online Education in the United States, 2005," The Sloan Consortium, Needham, MA 2005.
14. D. Mioduser, "Internet-in-education in Israel: Issues and trends," Educational Technology, Research and Development, vol. 49, pp. 74-83, 2001.
15. A. Shemla and R. Nachmias, "Current state of Web supported courses at Tel-Aviv University," International Journal of E-Learning, vol. 6, pp. 235-246, 2007.
16. I. E. Allen and J. Seaman, "Sizing the Opportunity: The Quality and Extent of Online Education in the United States, 2002 and 2003.," The Sloan Consortium, Needham, MA 2003.
17. A. Oern, R. Nachmias, D. Mioduser, and O. Lahav, "Learnet - a model for virtual learning communities in the World Wide Web," International journal of educational telecommunications, vol. 6, pp. 141-157, 2000.
18. C. J. Bonk and C. R. Graham, The handbook of blended learning: Global perspectives, local designs. San Francisco, CA: Pfeiffer Publishing, 2006.
19. R. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker, "Affective learning — a manifesto," BT Technology Journal, vol. 22, pp. 253-269, 2004.
20. R. B. Clariana, "Rate of activity completion by achievement, sex and report in computer-based instruction," Journal of Computing in Childhood Education, vol. 1, pp. 81-90, 1990.
21. D. Laurillard, "Computers and the emancipation of students: giving control to the learner," Instructional Science, vol. 16, pp. 3-18, 1987.
22. American Psychological Association, "Learner-centered psychological principles: a framework for school reform," 1997.
23. M. D. Williams, "A comprehensive review of learner-control: The role of learner characteristics," presented at Annual convention of the Association for Educational Communications and Technology, New Orleans, LA, 1993.
24. P. Zaharia, K. Vassilopoulou, and A. Poulymenakou, "Designing affective-oriented e-learning courses: An empirical study exploring quantitative relations between usability attributes and motivation to learn," presented at World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano, Switzerland, 2004.
25. J. B. Rotter, "Generalized expectancies for internal versus external control of reinforcement," Psychological Monographs: General and Applied, vol. 80, pp. 1-28, 1966.