Adaptive Online Learning Environment and Web Usage Mining

Péter Tóth
Obuda University, Trefort Ágoston Centre for Engineering Education
toth.peter@tmpk.uni-obuda.hu

Abstract— The realization of an adaptive online learning environment presupposes the full consideration of the existing knowledge of the individual and their learning characteristics both in the area of course as well as learning management. Therefore during the learning process diagnostic and formative control, evaluation and assessment, with web usage mining as an efficient means, has a special significance. Educational technology and multimedia, a course in technical teacher training, formed the basis of the study. By the application of web mining methods we were trying to find an answer to the question whether any conclusions from the patterns of online learning activities were to be drawn as to preferred learning characteristics, methods and strategies and also whether the major variables of online learning behavior were possible to define.

I. INTRODUCTION

The application of a virtual learning environment and computer based learning has become widespread in Hungarian higher education. [1] [2] Questions of quality and adaptivity are increasingly gaining dominance, manifested in course development and course management which takes the individual specialities of learners well into consideration. In order to increase the adaptivity of the learning process, we need to have exact and relevant information on the learner’s learning characteristics and preferred learning strategies in an online environment. In this we may be aided by web mining methods, which process data from interaction between the learner and the learning objects.

There are two kinds of procedure to be followed in selecting the appropriate online teaching strategy. The top-down method involves the identification – usually by an online questionnaire - of the learner’s learning style first, then, on the basis of this, we can conclude the preferred learning strategies and, from these, the learning methods, means and forms. The other possibility is the so-called bottom-up method, when we proceed just in the opposite direction by starting with the patterns of concrete learning activities – by using for example methods of web mining –, arrive at conclusions on learning characteristics and preferred learning methods, on the basis of which learning strategies are relatively easy to identify. Now to the learning strategies preferred by the individual already recommended methods of syllabus processing, learning routes, individual course management methods as well as tutorial methods may be assigned.

II. PERSONALIZED E-LEARNING AND WEB MINING

The distinction between web mining and data mining was made as early as in 1997, however, it only became a field of research in its own right over the last 10 years. There are two approaches to the interpretation of web mining. The process-oriented theory regards web mining the sequence of successive tasks [3], whereas the data-centered concept discusses different web mining methods according to the types of the web data analyzed. [4] It is rather the second approach that has become more accepted, according to which web mining is a special area of data mining, applied for analyzing data created on web servers, that is web content mining, web structure mining, web usage mining are to be discussed. A popular synonym for web mining is the expression ‘knowledge discovery in web databases’, too. The work of Kosala and Blockeel [5] presents an overview of research into this field up to the year 2000. The study by Srivastava and colleagues also deserves attention, investigating the behavior of users, web mining methods, which process data from interaction between the learner and the learning objects.

From the point of view of our research their endeavors made to identify web metrics and measurements (e.g. visits to pages, visit-purchase rate), to analyze click-streams describing the decision-making process (click-stream analysis) (e.g. the process between entering the web store and purchasing or in fact failing to do so) and to investigate the time factor of web communities, contents and structures are to be highlighted in their work.

Khribi and colleagues still look upon the personalization of virtual learning environment, that is the consideration of the learners’ special characteristics both in the course of planning and learning management as a problem unsolved. That is why they pose the question of an adaptive course management, by which they mean a dynamic restructuring of the course, the adaptive selection and personalized composition of the learning objects as well as an adaptive navigation support. Their adaptive e-learning flow-model can be divided into two phases. In the course of modeling (offline mode) at the formation of the e-learner profile they take into account information gained for example from the interaction between learner and learning environment (e.g. preferred learning objects, learning routes), the existing knowledge of the student, his learning characteristics and style. This is followed by the formation of homogenous groups of e-learners with similar learning characteristics through the application of cluster analysis and associative methods (typifying). In the phase of counseling (online mode) first the observation and analysis of the learning activities of the learner who is
just being active in the course takes place along the parameters mentioned in connection with modeling, then the learner is assigned to the group with the same learning characteristics as his. After this propositions concerning learning objects and learning routes are made through the application of filters focusing on syllabus content as well as collaborative activities. [7]

A similar model of the formation of a personalized electronic learning environment is to be seen at Jain and colleagues, too, who created an adaptive system by using means of the semantic web as well as web mining. In their model learning objects most suitable for the individual characteristics of learners are selected through the application of so-called personalized e-learning services. In their colleagues, too, who created an adaptive system by using filters focusing on syllabus content as well as collaborative activities. 

III. EXAMINED COURSE, METHODS, OBJECTIVES

It was the course implemented in a virtual learning environment (Moodle) of a subject taught in professional teacher training, educational technology and multimedia, that formed the object of the examination. Students taking part in correspondence courses learn the aspects, methods and means of the development of information media (overhead projector, video film, photograph, chart, animation, computer presentation, etc) and the use of educational technological aids applicable in the course of their pedagogical work (e.g. overhead projector, video projector, document camera, camera) within the frame of this subject.

An increase in the dynamically changing syllabus content and a decrease in contact lessons necessitated the development and later the application of electronic syllabus in this subject. As a result of the development a four-module (basics of educational technology, digital imaging, image editing, video editing) multimedia based interactive electronic syllabus was created, which, besides the introduction and application in pedagogy of education technological tools, drills the process of the development of information media, that is, the acquisition at a skill and proficiency level of editing programs is highlighted.

Congruence with this dual objective was also reflected in setting the electronic format syllabus content. For the acquisition of information photos, images, texts (written and narrative), animation, and video, while for the introduction of editing algorithms animation supported by narrative explanation and videos were integrated in the electronic syllabus.

Besides the electronic and interactive syllabus contents further objects that support studies were applied in the course, for example discussion boards, wiki and dynamic glossary. Along with that the knowledge of learners was measured through online tests and the solution of productive tasks.

The theoretical model of the educational technology and multimedia course is shown by Fig. 1, whereas its concrete interactive syllabus structure is represented by Table I.

In the course of the examination the CRISP-DM model known in data mining was applied, which is also well usable in analyzing through web mining the database formed in applications for educational purposes. [14]

It, however, must be taken into consideration that with respect to a particular course a relatively significant scale, repeated sequence of activities by a relatively small number of test persons is to be taken into account, which in turn makes the generalization of experience difficult.

From among the methods applied during the examination accounts provided by the statistical system of Moodle and in data mining the frequency, sequence and cluster analysis and that of prominent values, classification and associative procedure are to be highlighted.

Data mining procedures appropriate for analyzing online courses and results of their application are treated in several researchers’ works. [9] [10] [11] [12] [13]

In the course of the research SPSS (IBM) Modeler (Clementine) program and its Web mining node was used, which offers concrete algorithms to realize the data mining methods mentioned above.

For an interpretation of the results of the examination three basic concepts must be made clear. The learner interacts with the screens of the course and its learning objects with the purpose of learning (e.g. opens, downloads and uploads a page or a document, does a test, contributes to the discussion board). So the interaction is to be interpreted at the level of screen pages (php, html, xml), probably files or their larger units, that is events. Therefore analyses focusing on click-streams or learning events (e.g. submitting assignments, activity at forums, wiki-notes) can be distinguished. During a particular visit – from entering the course to leaving it – the learner opens several screen pages, gets into contact with several learning objects and performs a full sequence of operations. The examination directed at click-streams is called microanalysis whereas the one directed at learning events (objects) is called macroanalysis.

![Figure 1. The model of the educational technology and multimedia course](image-url)
All the operations by the learners are administered in the so-called logfile. During the processing, to these notes in the logs User ID, Visit ID and Event ID are assigned, which makes segmentation according to learners, visits and events possible. The date of the operations performed is also recorded, that is tracking the activity of learners in the course is easily solved.

On the basis of the succeedences of learning activity as well as of the frequency in time of the visits a notion may be formed of the learning methods and strategies. Also, a collation and typifying of visit habits can be realized.

Three variables were introduced for the frequency analysis of the online course. „Days Active“ means the time lapse between the learner’s first and last visit to the course, „recency“ denotes the number of days that passed since the last visit, while learning „frequency“ refers to the number of visits during the time interval under examination (Fig. 2).

Our examination belongs to the category of web usage mining since our objective is the analysis of visit structure, click-streams and learning activity as well as the identification of learner habits, methods and strategies.

Variables of online course frequency analysis

Figure 2. Variables of online course frequency analysis

Based on the above we looked for an answer to an open question. How are the syllabus processing procedures as conceived by the developer and as finally realized by the learners relate?

45 persons (28 men and 17 women), students of a correspondence course in engineering education participated in the research. Their specializations were as follows: 8 mechanical engineers, 16 electric engineers, 6 information technology specialist engineers, 6 light industry engineers and 9 technical managers. With the exception of 4 of them they all teach at vocational secondary schools. The average age of the group is 39.68 years with the youngest person being 27, while the oldest one 58 years old. 16 of them teach in Budapest, 12 in a city and 17 in small towns. 49.17% of the people in the experiment use an IT device in class regularly, 27.69% do usually portable computers and projectors. It is mainly in the preparation for the class (56.47%) and in communication (76.35%) that IT devices are used outside class. All the persons in the experiment own a computer at home with all but 3 of them with Internet access.

IV. Results

The answer to both questions is to be given by an analysis of the learning activities (segmentation of visits by learners or events as well as their temporal analysis).

Our examination may focus on either learning objects (to these learning events such as for example forum activity or submitting assignments) or screen pages (e.g. SCORM or HTML, XML base pages). The former is called macro- while the latter one microanalysis.

Through the examination we arrived at the conclusions below.

An important means of analyzing learning activities is segmentation according to event (learning object), visit (sequence of learning events) or visitor (learner). In both cases we wish to create homogeneous groups from the aspects of learning objectives and learning behavior.

The aim of visit-based segmentation is the identification of the particular patterns and clusters of learning activities, the comprehension of what and why a learner is doing during his activity in the course and finally to aid the developer in recognizing the strengths and weaknesses of the course and in formulating the developmental objectives founded on these.

Based on data characteristics of the use of the Moodle system (time of entry, length and frequency of connection, learning objects used, etc) a learning algorithm groups visits (sequence of learning events) or users (learners in our case). Segmentation takes place in two steps. In the phase of model construction visit segments are created by a so-called two-step clustering, then they are classified by C5.0 algorithm producing decision trees. In the second phase the individual visits are „given scores”, then saved in a data file which describes the segments of each visits.

The result is the cluster identified, which later may be used for the segmentation of learners, that is the formation of learner groups as well as for certain learning propensity analysis and syllabus-development counseling.

The results of the visit-based segmentation by the Advanced Visit Segmentation stream (focusing on the sequence of learning events) are presented below. The isolated clusters are as follows:

- **Independent learning directed at the acquisition of basic concepts** which mostly meant the processing of the interactive electronic syllabus of the basics of educational technology, the inspection and extension of the glossary as well as solving the related self-check tests. During their visits to this cluster learners often consulted their teachers with their problems or joined the discussions on the forum. (C1)

- **Learning directed at the acquisition of construction algorithms** which meant the acquisition of the basic operation sequences of editing and imaging programs by viewing or listening to animations and narrative explanations besides the traditional syllabus contents independent of time. Learners did the self-check tests and practice tasks related to the categories (reproductive knowledge). (C2)

- **Participation at moderated discussions** related to subject categories. The objective of these visits was exclusively the inspection or initiation of comments. (C3)
- The upload to the course of tasks accounting for the creative application of the material learnt then the inspection of the assessment. (C4)
- Checking the acquisition of information, the inspection of the assessment. (C5)

These clusters may also be interpreted as a significant group of the visits focusing on independent learning, while another one on the “justification” of learning achievement (submitting assignments, tests, comments on forums). According to another interpretation all but one of the activity groups represent the individual form of learning with the one exception representing a community form of it (C3). A mixed pattern of learning objects was less typical.

<table>
<thead>
<tr>
<th>TABLE II. THE CHARACTERISTICS OF CLUSTERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>C1 207</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>C2 348</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>C3 397</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>C4 467</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>C5 240</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

In Table II the data of each cluster is given. As can be seen the number of visits directed at processing interactive syllabuses (basics, editing algorithms) is relatively small, however, the average time devoted to learning is high, showing significant individual differences though. Visits directed at tests, participation at discussion forums and submitting assignments, in other words the other three clusters, show a picture diametrically opposed to this, that is the number of visits is relatively high, but the length of connection, with the exception of forum discussion, is short and its individual standard deviation lower.

The creation of homogeneous groups may take place according to learner behavior, too. One means of segmentation is the User Mode Determination stream. Segmentation according to users is similar to the visit-based one. With the use of activity successions the most characteristic learner groups can be isolated by a two-step cluster analysis, from which modes of behavior may be concluded:

- Interactive syllabus processing by preferred time-dependent media (viewing or listening to animations and narrative audios besides the traditional textual and image contents)
- Interactive syllabus processing by less preferred time-dependent media (ignoring animations and narrative explanations)
- Doing practice tests (related to digital imaging and digital image editing)
- Passive community learning (following forum discussions and wiki notes, reading tutorial messages)
- Active community learning (commenting on discussions, entering notes in wikis and glossaries)
- Checking independent work (doing online tests, uploading productive tasks to the course)
- Assessment of independent work (inspection of the results of online tests and tutorial reflections on assignments)

As we see, all this supports the results of event based segmentation.

As shown by the above, learners present significant individual differences in their course activity. Let us therefore examine the visits in the dimension of time, too.

In order to perform that another method of segmentation according to learners may help. Segmentation may take place in the dimension of recency, frequency and the individual learning interval (days active). The examination of these variables makes the understanding of online learner behavior easier. For instance learners who have entered the course recently are more likely to return there than those who only visited it some time ago or those learners who frequently visit the course will probably also do so later. The learner who often contributes to a forum discussion will more probably do so later, too.

The E-Channel User RFM Classifications stream is suitable for the identification of online learning behavior in the dimensions of recency – frequency as well as of recency – frequency – days active. Segmentation according to these variables is done by a division into five parts: high (80 percent, top 20%), medium-high, medium, medium-low and low (20 percent, bottom 20%).

![Figure 3. The segmentation of learner behavior according to frequency and recency](image-url)

Fig. 3 shows learner behavior in the course with regard to frequency and recency. The rarer learners enter the course the more likely that they made their last visit there a long time ago. Learners of average activity made their last visit only a short time ago. Quite a long time has passed since the last activity of online learners producing
a considerable number of visits, in other words these visits were restricted to a short period of time, for example to that before the online check tests or submitting assignments.

Consequently, in this dimension four types of learner behavior are to be distinguished.

- The learner visited the course only a few times and briefly and his activity was mostly deadline-related, like for example the submission of some assignment, or an online test, however, it was not proceeded by intensive learning on the course (casual online learner).
- Medium-frequency course visits in an even distribution during the term (average online learner).
- The learner visited the course several times but it was always deadline-related, like for example active learning before an online test („campaign learning”).
- A great number of course visits in an even distribution during the term. The learner entered the course several times a week, got informed on forum discussions, viewed a particular syllabus unit, etc (dedicated, loyal online learner).

Average visit frequency is presented in a cross table with regard to days active and recency (Table III). The highest value of days active was 98, the lowest one 64, while the average was 89 days.

The daily average of visit frequency is highest when days active are on the top and recency the lowest. At the same time, the majority of learners registered in the course or finished their activity by submitting the last task and they very often closely followed the course. The average was 89 days.

The daily average of visit frequency is highest when days active are on the top and recency the lowest. At the same time, the majority of learners registered in the course or finished their activity by submitting the last task and they very often closely followed the course. The average was 89 days.

The cells contains the mean values of daily visit frequency

### TABLE III.
THE SEGMENTATION OF LEARNER BEHAVIOR ACCORDING TO DAYS ACTIVE AND REGENCY

<table>
<thead>
<tr>
<th>Recency</th>
<th>Days active</th>
<th>Low</th>
<th>Medium-Low</th>
<th>Medium</th>
<th>Medium-High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.092</td>
<td>0</td>
</tr>
<tr>
<td>Medium-Low</td>
<td>0</td>
<td>0</td>
<td>0.024</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td>0</td>
<td>0.018</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medium-High</td>
<td>0.019</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

V. SUMMARY

Based on the examinations to the open questions posed at the beginning of the research, that is concerning online learning specialities, preferred strategies, as well as the comparison of learning routes as conceived by the developer (expert) and as finally realized by the beginners (learners), the following answers may be given:

A. With respect to the relation to the structure

By the application of frequency-, sequence- and cluster analysis four kinds of typical online learning strategies are to be distinguished on the basis of the relation to the structure, namely the conscious and uncertain followers of the structure, structure-abstractor and the unstructured scanner.

The consciousness and uncertainty of structure following is related to the cognitive stylistic characteristics of the field.

The **structure follower online learner** used learning objects mostly in the order given by the course developer (based on the results of the macroanalysis) and in processing the interactive electronic syllabus he consequently followed the route defined by the developer (based on the results of the microanalysis). He primarily focused on the acquisition of the syllabus content and not on getting familiar with the structure and the navigation tools.

The learner independent of the field manages structure with confidence and has mental models ready to find analogy with ease. However, the structure following of the field-dependent learner is characterized by uncertainty, has no mental models ready therefore spends longer time over the processing of a page. A syllabus allowing of multi-bifurcation fits the former type while that of a linear structure primarily fits the latter most.

The **structure-abstractor learner** opened almost all of the learning objects of the course, took the things to be acquired into account and in processing the electronic syllabus contents tried all navigation and bifurcation possibilities. First he „acquired” the structure and only then did he focus on the information to be acquired.

The **unstructured scanner learner** did not visit the learning objects in the order designed by the course developer, his navigation behavior is random-like, often proceeded towards contents which seemed more spectacular and usually only concentrated on tasks with a deadline. He visited the course rarely and at uneven intervals. This learner needs a strictly set syllabus structure and permanent tutorial attention.

The orientation of abstractor and scanner learners is largely aided by the placement of positioning elements in the structure.

B. With respect to the timing of learning

The macroanalysis of learning activities may also be performed according to timing (frequency analysis). For this were learning frequency, period and days active introduced. On this basis four types of learning activity were distinguished: frequent – even (dedicated, loyal online learner), frequent – uneven („campaign learning”), occasional – even (average online learner), rare – uneven (superficial online learner).

Comparing the results of the two kinds of examination we got the patterns seen in Table IV.

Although to a different extent, a certain learner „dropping off”, the giving up of learning due to cognitive or emotional-volitional reasons, is continuously typical of the processing of syllabus units. Regarding the syllabus units (the result of the macroanalysis) it was mostly typical of the ones processed first, while regarding the screen contents (the result of the microanalysis) it was typical of the first few pages. The abortion of syllabus processing was mostly characteristic of unstructured scanners but also to a certain extent to structure abstractor
learners. The latter naturally returned later to continue the syllabus acquisition.

### TABLE IV.
THE PATTERNS OF ONLINE LEARNING STRATEGIES

<table>
<thead>
<tr>
<th>Learning strategies</th>
<th>The timing of learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequent – Even</td>
</tr>
<tr>
<td>Conscious structure follower</td>
<td>x</td>
</tr>
<tr>
<td>Uncertain structure follower</td>
<td>x</td>
</tr>
<tr>
<td>Structure-abstractor</td>
<td>x</td>
</tr>
<tr>
<td>Unstructured scanner</td>
<td>x</td>
</tr>
</tbody>
</table>

C. With respect to subject areas

Based on the relation to particular subject areas holistic and atomistic learning behaviors are to be distinguished. The holistic learner processed all the learning objects belonging to the given area one after the other in the course of the same visit, while his atomistic colleague’s visits mostly focused on one or two objects, in other words, his learning composed of a stream of parts of processes.

D. With respect to syllabus-units

According to the length of successive page streams enduring, easily giving up and testing learner behaviors may be distinguished. The first one is characterized by a continuous processing of the entire syllabus unit, the second one by the acquisition of certain parts only, and the third one by inspection which is trying to orientate himself and gather information. While the second one only rarely does so, the third one almost always returns to processing the syllabus unit.

REFERENCES


