E-learning and Web Mining

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

Abstract — The web mining is a very effective data mining approach in the internet-based segments of the business world, appropriate equipment to discover useful information or knowledge from the web hyperlink structure, page content and usage log. We use this method for interpretation and understanding of students’ activities in an e-course and for development the course materials and learning activities as well. Here we are going to present the results exposed by quality assurance in connection with the students’ learning activity, the structure of the syllabus as well as the navigational opportunities. By analysis of student behavior we receive some beneficial information for course development and learning management.

I. COURSE EVALUATION AND BASICS OF WEB MINING

The ADDIE model is a systematic instructional design model consisting of five phases: Analysis, Design, Development, Implementation, and Evaluation. Various versions of the ADDIE model are applied in a wide area of training development. [1] Instructional theories also play a determinative function in the shaping of instructional materials. Different theories such as behaviorism, constructivism, social learning and cognitivism help in design and preparation the outcome of instructional materials.

Concentrating on the evaluation phase, Kirkpatrick's Four Level Evaluation Model may be considered the best known training methodology. The four levels of evaluation consist of student reactions, learning outcomes (knowledge, skills, attitude), student behavior (formally – testing, informally – observation) and final results or impacts in daily work. [2]

Web mining methods concentrate on techniques that could predict students’ behavior while they are interacting with the given virtual course. Its goals are to discover any meaningful patterns from data generated by client-server transactions on one or more Web localities. These methods could be categorized into two sub-types, such as learning a student profile or learner modeling in adaptive interfaces and learning student navigational patterns. Web usage mining is very suited for personalizing web experience for a learner, Virtual Learning Environment (VLE) site and course development, usage characterization. [3] [4] By way of these methods developers could improve the effectiveness of the virtual courses by adapting design or by directing the students’ behavior towards satisfying the learning objectives of the course.

Web mining, was first mentioned by Etzioni [5], who suggested that traditional data mining techniques for finding hidden patterns in huge databases, can be used to web-based information. Web mining is an emerging method in education research, assisting instructors and developers in improving learning environments and supporting decision-making of policymakers. [6]

Models for applying usage mining as a research method in education were suggested by Pahl [7] and Zaïane [8], although earlier research already discussed the potential of analyzing on-line courses using this method. 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. [7]

According to Liu data mining is also called knowledge discovery in databases. It is commonly defined as the process of discovering useful patterns or knowledge from data sources, e.g. databases, texts, the web, etc. The patterns must be valid, potentially useful and understandable. “Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization. Web mining aims to discover useful information or knowledge from the web hyperlink structure, page content and usage log. Based on the primary kind of data used in the mining process, web mining tasks are categorized into three main types: Web content mining, Web structure mining, and Web usage mining.” [9]

From point of view of virtual courses Web content mining is the process of extracting useful knowledge from the contents of learning objects (text, image, audio, video). Content data corresponds to the collection of facts a learning object was designed to transmit to the learners.

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the given virtual course. It can be further divided into two sub-types based on the kind of structure information used: hyperlinks, document structure. These methods by means of students’ visits deal with the analysis of visit structures, click series and with the planning strategy of web documents.
The application of web mining in VLE is an iterative cycle in which the excavated knowledge should "enter the loop of the system and guide, facilitate and enhance learning as a whole, not only turning data into knowledge, but also filtering mined knowledge for decision making". [12]

The CRISP-DM (CRoss Industry Standard Process for Data Mining) as a well-known process model [13] consists of six phases (Fig. 1): learning and data understanding, data preparation, modeling, evaluation, deployment.

II. INTRODUCTION OF ANALYZED COURSE

The most technology-demanding subject in technical teacher training is Educational Technology and Multimedia. In the framework of this subject, about 60 students on full-time and correspondent courses learn how to use education technological devices in their pedagogical work (overhead projector, video, video projector, documentary camera, etc) as well as the means and equipment of preparing and developing information media (transparency, video film, photograph, figure, digital presentation, etc). [11]

The increase in dynamically changing syllabus content and the decrease in the number of contact classes made it necessary to develop, then use electronic syllabuses in this subject first. As a result of the first development we composed a multimedia base syllabus of four modules (basic skills, digital image editing, digital vector graphics editing, digital video editing), which we made available to our students on an optical disc. The electronic syllabus does not only show the structure of education technological devices, their installation and the possibilities of their application in education, but it also surveys and drills the development processes, strategies and learning characteristics from the syllabus relations, that is the learning activity or, generally, the student – virtual learning environment, in other words, the student – Moodle objects, and the student – electronic syllabus sets two tasks to the student, namely covering the syllabus units in the nodes.

Meeting this dual requirement was reflected in setting the electronic syllabus contents. In the course of communicating information, images, illustrations, texts (written and narrative), animations and videos, while in the course of presenting structural algorithms, animations supported by narrative explanation were applied.

Processing the syllabus was conceived in a blended form. Students took part in 3-hour practices per week, where they learnt how to handle education technological devices and how to use, at a basic level, all those programs which make the preparation of the most frequently used visual aids, teaching aids and electronic syllabuses in technical and adult training possible.

Between guided practices students individually had to process electronic syllabuses, prepare their homework, and take part in forums. Through the opportunities offered by wiki, joint syllabus development in a selected topic also took place. Student skills were measured by using the Moodle test module.

As investigation was partly targeted at the processing specialties of the syllabus inserted in html format, let us first see the structure of that. The electronic syllabus was divided into three modules (Basic Skills, Digital vector graphics editing, Digital image editing) and then 2 to 3 further sub-modules were separated within each module. One sub-module covered 2 or 3 units, 12 to 22 mixed-structure (hierarchical-linear) display pages on average.

In the Educational Technology and Multimedia course (Fig. 2) we wanted to show the sequence of syllabus units by the order we placed them in the course as well as by numbering them. At the same time, the student was free to decide the order in which he wanted to process the syllabus.

We preferred subsections within a syllabus unit, and linear navigation within those, although we also made sequential navigation possible after the main page. We did not design linear navigation only one-way, but we guaranteed stepping back and exit from Help, too.

Self check tests were also built in the Moodle course, which students could use any time they wanted to test their knowledge. These tests could be integrated into their learning process at any stage.

From all this it follows that within the virtual learning environment the designer of the course and the electronic syllabus sets two tasks to the student, namely covering the route as he planned it and understanding as well as accessing the syllabus units in the nodes.

Conclusions may be drawn about certain cognitive processes, strategies and learning characteristics from the student – virtual learning environment, in other words, the student – Moodle objects, and the student – electronic syllabus relations, that is the learning activity or, generally, behavior of the students may be open to investigation. Peculiarities can be explored and certain learning habits can be distinguished and typified. Two cognitive maps take shape from the context of syllabus developer – electronic syllabus – student. More precisely saying, two cognitive maps are compared. The map conceived and created by the developer (Fig. 2) on the one hand, and the cognitive map as finally realized by the student. We may come to conclusions about certain learning characteristics by comparing these. The simpler a cognitive map is, the simpler its inner representation, and the more complex it is, the longer time it

In the meantime, emphasis shifted increasingly from a multimedia base individual learning environment to collaborative and cooperative learning environments. Therefore we also revised our electronic syllabuses to meet the requirements of the virtual learning environment. Since these syllabuses were available in a html format, it seemed practical to insert them in the same format in Moodle. To the electronic syllabus organized into modules several Moodle objects were added (forum, submission, test, wiki). [11]

The most technology-demanding subject in technical teacher training is Educational Technology and Multimedia. Therefore we also revised our electronic syllabuses to meet the requirements of the virtual learning environment. Since these syllabuses were available in a html format, it seemed practical to insert them in the same format in Moodle. To the electronic syllabus organized into modules several Moodle objects were added (forum, submission, test, wiki). [11]

The CRISP-DM (CRoss Industry Standard Process for Data Mining) as a well-known process model [13] consists of six phases (Fig. 1): learning and data understanding, data preparation, modeling, evaluation, deployment.

II. INTRODUCTION OF ANALYZED COURSE

The most technology-demanding subject in technical teacher training is Educational Technology and Multimedia. In the framework of this subject, about 60 students on full-time and correspondent courses learn how to use education technological devices in their pedagogical work (overhead projector, video, video projector, documentary camera, etc) as well as the means and equipment of preparing and developing information media (transparency, video film, photograph, figure, digital presentation, etc). [11]

The increase in dynamically changing syllabus content and the decrease in the number of contact classes made it necessary to develop, then use electronic syllabuses in this subject first. As a result of the first development we composed a multimedia base syllabus of four modules (basic skills, digital image editing, digital vector graphics editing, digital video editing), which we made available to our students on an optical disc. The electronic syllabus does not only show the structure of education technological devices, their installation and the possibilities of their application in education, but it also surveys and drills the development processes, strategies and learning characteristics from the syllabus relations, that is the learning activity or, generally, the student – virtual learning environment, in other words, the student – Moodle objects, and the student – electronic syllabus sets two tasks to the student, namely covering the syllabus units in the nodes.

Meeting this dual requirement was reflected in setting the electronic syllabus contents. In the course of communicating information, images, illustrations, texts (written and narrative), animations and videos, while in the course of presenting structural algorithms, animations supported by narrative explanation were applied.

Processing the syllabus was conceived in a blended form. Students took part in 3-hour practices per week, where they learnt how to handle education technological devices and how to use, at a basic level, all those programs which make the preparation of the most frequently used visual aids, teaching aids and electronic syllabuses in technical and adult training possible.

Between guided practices students individually had to process electronic syllabuses, prepare their homework, and take part in forums. Through the opportunities offered by wiki, joint syllabus development in a selected topic also took place. Student skills were measured by using the Moodle test module.

As investigation was partly targeted at the processing specialties of the syllabus inserted in html format, let us first see the structure of that. The electronic syllabus was divided into three modules (Basic Skills, Digital vector graphics editing, Digital image editing) and then 2 to 3 further sub-modules were separated within each module. One sub-module covered 2 or 3 units, 12 to 22 mixed-structure (hierarchical-linear) display pages on average.

In the Educational Technology and Multimedia course (Fig. 2) we wanted to show the sequence of syllabus units by the order we placed them in the course as well as by numbering them. At the same time, the student was free to decide the order in which he wanted to process the syllabus.

We preferred subsections within a syllabus unit, and linear navigation within those, although we also made sequential navigation possible after the main page. We did not design linear navigation only one-way, but we guaranteed stepping back and exit from Help, too.

Self check tests were also built in the Moodle course, which students could use any time they wanted to test their knowledge. These tests could be integrated into their learning process at any stage.

From all this it follows that within the virtual learning environment the designer of the course and the electronic syllabus sets two tasks to the student, namely covering the route as he planned it and understanding as well as accessing the syllabus units in the nodes.

Conclusions may be drawn about certain cognitive processes, strategies and learning characteristics from the student – virtual learning environment, in other words, the student – Moodle objects, and the student – electronic syllabus relations, that is the learning activity or, generally, behavior of the students may be open to investigation. Peculiarities can be explored and certain learning habits can be distinguished and typified. Two cognitive maps take shape from the context of syllabus developer – electronic syllabus – student. More precisely saying, two cognitive maps are compared. The map conceived and created by the developer (Fig. 2) on the one hand, and the cognitive map as finally realized by the student. We may come to conclusions about certain learning characteristics by comparing these. The simpler a cognitive map is, the simpler its inner representation, and the more complex it is, the longer time it
takes to understand and note. The cognitive network which is as simple as possible and repeated at each syllabus unit needs the least possible attention from the student during navigation, so emphasis falls on the acquisition of information in the nodes. [11]

Having said all that, answers are now being sought to the following two questions by applying methods of web mining:
- To what extent did students in the Education Technology and Multimedia course prefer the order of processing syllabus units as suggested by the developer?
- How did the learning process conceived by the developer relate to the one realized by the students within a syllabus unit?

III. WEB MINING ALGORITHMS

The main objective of the investigation was to explore by web mining methods the most important characteristics of student behaviour or, in other words, the learning activity during the accomplishment of the Moodle courses. The results presented in the previous chapter do not describe classic e-Learning-base distant learning but blended-form full time training where traditional classroom education is specially integrated with online and offline learning methods.

In introducing learning activity two approaches were followed. On the one hand, the role in the students’ learning process of the objects which produce learning activity on the virtual course was explored (macroanalysis), and on the other hand the microstructure of processing electronic syllabuses was investigated (microanalysis).

So-called offline web mining methods such as Google Analytics and SPSS Web mining for Clementine and its web mining node are able to provide opportunity for a lot more profound, more comprehensive and more scientific analyses, far beyond descriptive statistics. It is an aggravating circumstance for the application of online tools that within the virtual learning environment several simultaneous courses are running. The isolation of related results as well as their survey are often too complicated. It is also difficult to perform analyses concerning the levels of the objects comprising the module and the pages comprising SCORM module and the html base syllabus. Clementine is also capable of extracting related data from the log file.

A. Macroanalysis by SPSS Clementine

In the course of macroanalysis we wanted to see what role Moodle objects inserted in the course, html base electronic syllabus units, static glossaries, submission, forum and self check tests played in online and offline learning. To answer this set of questions the visit level analysing streams of the programme were used. First, the electronic syllabus was examined, which is divided into three modules (basic skills, digital image editing, digital vector graphics editing). The modules consist of syllabus units (separate objects), and those in turn consist of subsections accessible from the menu. The pages of the subsection are on the one hand linearly connected and, on the other hand, their first page is accessible from the menu. In the case of macroanalysis the succession of modules and syllabus units, while in the case of microanalysis that of subsections and pages during visits were examined.

<table>
<thead>
<tr>
<th>Item</th>
<th>% Starting Activity</th>
<th>Visits</th>
<th>Dropout Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital image editing 1</td>
<td>100.00</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>Digital image editing 2</td>
<td>35.22</td>
<td>56</td>
<td>64.78</td>
</tr>
<tr>
<td>Digital image editing 3</td>
<td>17.61</td>
<td>28</td>
<td>50.00</td>
</tr>
<tr>
<td>Self-check Test</td>
<td>8.3</td>
<td>13</td>
<td>53.57</td>
</tr>
</tbody>
</table>

Fig. 3. The Succession of Sub-modules Comprising Visits – Digital Image Editing

Although all modules and syllabus units were visible simultaneously on the course students followed the order set by the teacher and the way syllabus processing took place during practices. Midterm tasks and module-end tests also prevented "campaign-like" learning at the end of the term. In processing syllabus units (usually 2 or 3) comprising the modules, two types of learning strategies could be distinguished. The most dominant form of learning strategy was the one when a student visited aimed at processing more than one (usually two) syllabus units, but aiming to acquire only a single unit was also typical. There were few visits seen which aimed at processing all the syllabus units of a module besides doing the self check test (Fig.3). Thus the developer’s intention that students should learn not only before the check test appears to be realized. Although there is no real continuous student activity on the course, significant steps could be taken in the direction of collaborative learning by a well-structured course and well directed student activities.

Fig. 4. The Result of TwoStep Clustering with Identified Clusters

The above is reinforced by the so-called two-step cluster analysis, since it made the distinction of two learning strategies possible: the one limited to a Moodle object and the other, "successive" syllabus processing (Fig. 4).
Glossary was accessible in two ways: as an independent object from the course on the one hand and, on the other hand, through an icon in any of the pages of the syllabus unit. It is to be seen that students used this tool on a few occasions only, and even then mostly in the first modules only, so the Glossary has not become an integral part of the learning process.

The solution of self check tests mostly took place after finishing the particular module, usually within a single visit.

Activity in forums also turned out to be quite moderate. Mostly, it was functioning as a tutorial message board, that is students quite often paid a visit but rarely took the floor.

During the planning and organization of the course the above objects ought to be better integrated into the learning process. The following steps of organizing learning and course are recommended to be taken:

- dynamic Glossaries are to be created where students may have an active role, specifically by connecting the lexicon editing activity with the electronic syllabus
- the tutor is to encourage Forum activity and provide concrete instructions – or even links during the technical realization – referring to ideas and questions for the given syllabus and task setting, and he is to make students’ Forum activity a part of their assessment [14]
- mutual assessment is to be made possible in Submission
- SCORM base development makes the integration of self check tests into the syllabus possible.

As a result of the above, the number of students’ “flash visits” is decreased and a learning process takes shape which is more extensive in time and more diverse in content and makes students more active.

The student
- had contact with the virtual learning environment a few times but in a relatively even distribution during the term
- was a frequent visitor to the course in a relatively even distribution
- was a moderately frequent visitor to the course in a quite uneven distribution
- was a rare visitor to the course in an uneven distribution
- was a moderately frequent visitor to the course in an uneven distribution.

Days active are in inverse proportion to “recency”.

B. Macroanalysis by Google Analytics

Online web mining methods such as Google Analytics offer fast, easy-to-survey and clear answers to some of the above questions, thus satisfying the needs of everyday syllabus developers. The precondition of the application of these programmes is that the so-called “Follow” code (Javascript code) must be placed on the web-site to be analysed. This method does not offer the possibility to analyze data created before the installation of the code.

Figure 6 shows further data for describing students’ behaviour: pageview (a), spending time on the given VLE course (b), length of visit (c), depth of visit (d), loyalty of students (e).

![Fig. 6. Comparison of students’ behavior – learning process](image)

A pageview may be defined as a view of a page in the course. If a visitor hits reload after reaching the page, this will be counted as an additional pageview. If a user navigates to a different page and then returns to the original page, a second pageview will be recorded as well. Time on site is one way of measuring visit quality. If students spend a long time visiting the course, they may be interacting extensively with it. Length and depth of visit is a measure of visit quality. A large number of lengthy visits and high pageviews per visit suggests that students interact more extensively with the VLE course. Loyal students are frequently highly engaged with the brand and a high number of multiple visits indicates
good student retention. A high number of new students indicates strong student recruitment.

Figure 6 shows the comparison of students’ behaviour in three dimensions: pageviews week by week (a), pageviews hour by hour (b), average time on site (c). The following statements are to be drawn after the evaluation of the diagrams.

- Students’ learning activities are cyclical. They are very active before deadline of the learning tasks and at the end of the semester but between two “constraints” the students are prone to “relaxation.” Taking part in discussions in student forums and in common projects are appropriate methods to persuade them to work continuously.

- Average time on site is relatively smooth that indicates the students more often visit the course before learning task and spend less time there than during other period of the learning process when they go there more infrequently but pass more time.

Further indicators of the learning quality are length, depth and number of the visit and the recentness. The last parameter expresses the frequency with which learners return to the given course and it can indicate their level of engagement with the learning materials and their readiness to study.

C. Microanalysis by SPSS Clementine

As mentioned earlier, the microanalysis of the learning process focuses on the subsections comprising syllabus units as well as on the page which forms the syllabus. Those streams of the programme are usable well which aim at analysing pages and their sequence.

Although to a different extent, syllabus processing within a syllabus unit is constantly characterized by student “dropout”. With regard to syllabus units it was mainly typical of the ones processed first, whereas with regard to pages it was typical of the first 1 to 3 pages. Obviously, both are seen to originate in online learning and familiarity with the electronic syllabus. “Dropout” may also be interpreted in the dimension of the theoretical and the practical syllabus. Although it is more significant in the case of the former than the latter, it is constantly present in both. Students thus interrupting syllabus processing certainly return later and continue learning.

According to the length of successive series of pages, “steady” (“unflagging”), “giving up” (“flagging”) and “tasting” (“trying”) student behaviours can be distinguished. Having once entered a syllabus unit, the first one will typically learn the whole unit up to its end (Fig. 7). The student who easily “gives up” will not fully cover all the pages of the particular syllabus unit (Fig. 8) and he will escape when he becomes bored with the task which needs persistence. He is not certain to re-enter later and continue learning where he has left off. The “tasting” type will “leaf through” the material, get oriented and then interrupt the visit, but later he will return and fully deal with it.

<table>
<thead>
<tr>
<th>Page</th>
<th>% of Total Visits That Started Activity</th>
<th>Visits % of Visit Started</th>
<th>Dropoff %</th>
</tr>
</thead>
<tbody>
<tr>
<td>p7</td>
<td>26</td>
<td>52.83</td>
<td>6.67</td>
</tr>
<tr>
<td>p8</td>
<td>26</td>
<td>52.83</td>
<td>6.67</td>
</tr>
<tr>
<td>p9</td>
<td>20</td>
<td>52.83</td>
<td>6.67</td>
</tr>
</tbody>
</table>

To finish, we mention the last question of microanalysis, which aimed at the role of time-dependent media in syllabus processing. Average time allotted to process pages containing only texts or graphics was acceptable, but leafing through was frequent, too. It was mostly typical of “tasting” (“trying”) students. It is considered as leafing through when the student spends considerably less time over a given page than would be necessary to fully understand it, usually a few seconds only. Digital videos were usually played, especially when playing started automatically on opening the page, but this way it precedes reading the introductory text. It was not typical to interrupt playing automatic videos.

Playing narrative audio did not unfortunately become an organic part of online learning, which may be due to the “start playing” icon not being highlighted.

The above unambiguously shows that the realization of a really efficient collaborative (that is individual learning within a community) learning environment requires an approach different from the one seen in the case of the course under investigation from a constructional, organizational, tool-centred or tutorial point of view.

From a constructional and organizational aspect the virtual course to be processed in the form of blended learning is advisable to be divided into weeks. It is practical to assign to these weeks the syllabuses to be processed as well as all Moodle tools which aid processing. The structure of each week is identical. See a possible structure below:

- setting aims and requirements
- the (preferably SCORM base) virtual publishing of the syllabus
- defining and uploading student tasks
- cooperative work (online dialogue, wiki, workshop, etc.)
- online testing and assessment

Using tools primarily manifests itself in such first and second generation web-techniques which are characterized by content share and exchange, online dialogue and collaboration.
From the tutorial point of view the personal and continuous checking of the learning activity is to be emphasized besides the timing and encouragement of work on the course, the initiation of considerable activity in forums and debates, as well as prompt feedback on all student activity.

CONCLUSIONS

Our course evaluation model is based on the systematic instructional design model (ADDIE) and Kirkpatrick’s general evaluation model. This may be the basis of a modern quality assurance theory in course development and learning management. By web mining methods it is possible to analyse student behaviour informally so these tools may become effective elements of quality assurance. In introducing learning activity two approaches were presented by some samples. On the one hand, the role in the students’ learning process of the objects which produce learning activity on the Moodle course was explored (macroanalysis), and on the other hand the microstructure of processing electronic syllabuses was investigated (microanalysis).

The following conclusions are to be drawn after the evaluation of the results.

- Within a particular module students do not prefer the complete learning cycle that is the subsequent processing of syllabus units and the completion of the self-check test. During one visit, the processing of one or less frequently two syllabuses was dominant.
- The so-called reverse learning cycle that is the preference of self-check tests to information imparting parts was not typical.
- Within a particular syllabus unit, mainly with respect to pp 1-2, a significant “dropping off” was seen, which is to be interpreted as students’ orientation preceding actual learning.

Taking the above aspects into consideration, the directions of the electronic learning environment upgrade may be formulated as follows.

- It is practical to maximise the syllabus sub-modules in 14-15 screen pages.
- It is practical to incorporate self-check tests, similarly to the glossary, in the electronic syllabus, particularly at the end of the syllabus unit, the accomplishment of which is the prerequisite for opening the following syllabus unit.
- During development, a page finder window was also placed at the bottom of each page. This interfered with the sequential processing of the syllabus and resulted several times in the omission of substantial pages. A little more detailed exposition of the first, i.e. the introductory page makes it redundant to break the sequence.

REFERENCES