

Quality Assurance in Learning Management by Web Mining Usage

Imre J. Rudas

Óbuda University, John von Neumann Faculty of Informatics
rudas@uni-obuda.hu

Péter Tóth

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

***Abstract:** Web mining aims to discover useful information or knowledge from the web hyperlink structure, page content and usage log. The “user-centred” philosophy of this tool is in perfect harmony with the concepts of modern marketing, ergonomics, and learning management. This new approach, as opposed to the traditional “page-centred” philosophy, puts the users’ goals and intentions to the centre, and designs the services of the system accordingly. About 60 learners of technical teacher training took part simultaneously in processing the Educational technology and multimedia course, all their activities performed in Moodle learning environment were registered in a log file by the server. The processing of this log file was performed by the SPSS Web Mining for Clementine programme. Here we are going to present the first 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 behaviour we receive some beneficial information for course development and learning management.*

***Keywords:** virtual learning environment; student behaviour; web structure and usage mining*

1 Quality Assurance and Course Evaluation

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 behaviour (formally – testing, informally – observation) and final results or impacts in daily work. [2]

Web mining methods concentrate on techniques that could predict students' behaviour 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 modelling 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' behaviour towards satisfying the learning objectives of the course.

2 Philosophy of Web Mining

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 Zaiiane [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]

Based on the primary kind of data used in the mining process, web mining tasks are categorized into three main types:

- Web structure mining deals with the analysis of visit structures, click series, and with the planning strategy of web sites.
- Web content mining is an intelligent search agent, information filtering and categorization, and web query systems can be rated to the range of web content mining.
- The object of Web usage mining is the analysis of flow, the examination of customers' habits, and the increasing of interactivity.

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.

3 Basic Concepts

The log files from our server provide the data that we need for web mining. A web mining program reads and processes the web server log files, collecting information that we can then use for web analytics in general. These are among the data points extracted: hostname of the user; path to and name of the resource requested from the server, such as an HTML file; date and time that the request was made; authenticated user name for the user; referral page (the page that the user visited immediately before the current page); user identifier, also known as a user cookie. [10]

The Internet works via a request-response protocol. Each time a user looks at a web page with his web browser, the browser is making a request or a set of requests to a server. The web server sends back a response consisting of data, which the browser then interprets and displays. With each request it receives, the server writes an entry to its log file. Each request corresponds to a resource (a file)

on the web site, such as a home page. If the requested page contains additional resources, such as a narrative sound file, a video, an animation or an image, each additional item generates a separate request and, thus, a separate entry in the log file.

Web servers use a variety of log file formats. A basic log file entry might look something like this:

```
90.111.28.145 -- [15/Apr/2009:14:38:25 +0200] "GET /moodle/mod/resource/icon.gif HTTP/1.1" 304 - "http://mpi.banki.hu/moodle/course/view.php?id=37"
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322)"
```

IP address -- [date : time] "GET resource, protocol" file size - „URL address of linked web page" „browser type (and other data)"

Users, in this case students are the people who visit (request content from) our virtual course. A set of rules to our web log data is applied to determine whether a user who made a particular request is the same user as made other requests. The user-cookie method is the preferred method for identifying users at most web sites. It is transparent to most users, does not require them to log in, and produces the most accurate data for web analytics.

A *visit* is a series of requests to a web site made by the same user. More precisely, within the series, each request must occur no more than 30 minutes after the previous one. [10]

Events are the key interactions that a user makes with our web site, in this case in our virtual course. Defining and tracking events help us to understand how well our web site is fulfilling its intended purpose. We are able to focus less on the individual pages that users view and more on the actions that users perform. To define events, we must identify a set of pages (URLs) that interest us most for analysis purposes. These pages correspond to our events. A typical set of events in VLEs might include forum, submission, test, interactive learning material, etc. These can be given in event definition file (Table 1).

Table 1
Event Definition File

| Event category | Event name | Event Definition | Event Attributes |
|-------------------------------|-----------------------------------|--|---------------------|
| Forum | Forum | /moodle/mod/forum/index.php?id=37 | Forum_ID [Forum] |
| Interactive learning material | Digital vector graphics editing 1 | /moodle/file.php/37/Idofuggetlen/Da1/* | |
| Interactive learning material | Digital vector graphics editing 2 | /moodle/file.php/37/Idofuggetlen/Da2/* | |

4 Process of Web Mining Analysis

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:

- Learning understanding: This phase concentrates on understanding the learning objectives and requirements from a pedagogical perspective, and then converting this knowledge into a web mining problem definition, and a preliminary plan designed to achieve the objectives.
- Data understanding: This phase starts with collecting Moodle usage data and proceeds with activities in order to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- Data preparation: The data is cleaned and transformed into an appropriate format for web analytics.
- Modelling: The web mining algorithms are applied to build and execute the model that discovers and summarizes the knowledge of interest for the teacher or developer.
- Evaluation: The results or model obtained are interpreted and used by the tutor for understanding students’ behaviour and by the course developer during an improvement process.
- Deployment: Depending on the requirements, this optional phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

5 Introduction of Analysed 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 process of information media. In other words, a lot of emphasis is laid, besides communicating information, on the skilled acquisition of editing programs. [11]

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.

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]

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.

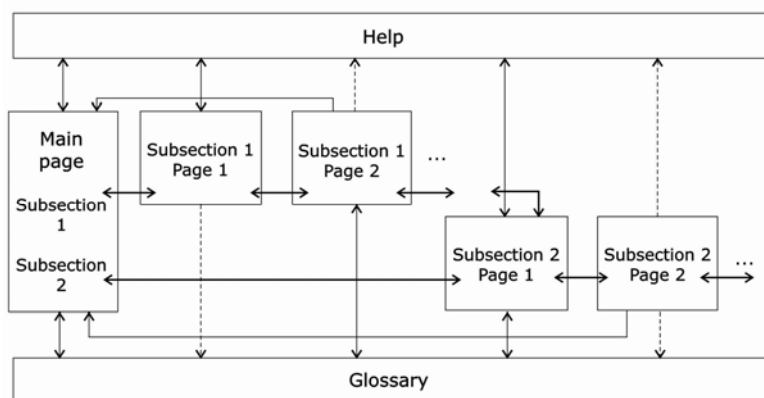


Figure 1
The Structure of Sub-modul

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 specialities 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 (Fig. 1), 12 to 22 mixed-structure (hierarchical-linear) display pages on average.

In the Educational Technology and Multimedia course 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, behaviour 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. 1) 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?

6 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-based 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 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.

6.1 Macroanalysis by Eventstream Visualization Stream

The Clementine contains two types of streams for macroanalysis. One of them analyzes a specific aspect of virtual course usage and another provides a general understanding of user, in this case learner behavior and course activity. Virtual courses commonly have one or more multi-step processes that learners are required to follow in order to achieve a specific goal. To learn the content of a

sub-module online, learners are typically required to progress through a series of steps from sub-module 1 to self-evaluation test. [14]

By 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 analyzing streams of the program were used. First, the electronic syllabus was examined, which is divided into above mentioned three modules. The modules consist of sub-modules (separate objects), and those in turn consist of units accessible from the menu. The pages of the units 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 sub-modules, while in the case of microanalysis that of units and pages during visits were examined.

| Item | % Starting Activity | Visits | Dropout Rate |
|-------------------------|---------------------|--------|--------------|
| Digital image editing 1 | 100.00 | 159 | |
| Digital image editing 2 | 35.22 | 56 | 64.78 |
| Digital image editing 3 | 17.61 | 28 | 50.00 |
| Self-checkTest | 8.1 | 13 | 53.57 |

Figure 2

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

The Eventstream Visualizaton stream looks at the traffic that is performing a specific sequence of activity in a known order. This type of analysis is particularly useful for virtual course that have well-defined, multi-step paths that the learner must follow to complete a specific action. The classic example of this is the path in a given sub-module: sub-module 1 – sub-module 2 – sub-module 3 – self-evaluation test. [14]

This report in Fig. 2 is designed to show the number of visits in which students follow the defined activity path. It is important to note that activities are noncontiguous. In other words, the student does not have to initiate the activities consecutively. As long as they occur in sequence and in the same visit, the visit qualifies as a valid occurrence.

In the example above, 159 visits include a visit the Digital image editing 1 sub-module. About 35% of these visits witness the student continuing to the Digital image editing 2, and 18% to the Digital image editing 3 sub-module. Of those, only 13 visits see the student proceed to the download demo activity. The learner completes the sequence in more than 8% of visits.

| Item | % Starting Activity | Visits | Dropout Rate |
|-----------------------------------|---------------------|--------|--------------|
| Digital vector graphics editing 1 | 100,00 | 86 | |
| Digital vector graphics editing 2 | 48,84 | 42 | 51,16 |
| Digital vector graphics editing 3 | 30,23 | 26 | 38,10 |
| Self-checkTest | 9,3 | 8 | 69,23 |

Figure 3

The Succession of Sub-modules Comprising Visits – Digital Vector Graphics 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 visit 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 (Figs. 2 and 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.

6.2 Microanalysis by Visit Page Funnels Stream

The microanalysis focuses on analysis at the page level rather than at the events level. The streams present some general summary statistics about our virtual course, e.g. which learners are using to access the course, which pages are getting the most hits, and what are the most commonly viewed sequences of pages (clickstreams). [14]

The analysis by Visit Page Funnels stream looks at the traffic that is performing a specific sequence of activities in a known order. This type of analysis is particularly useful for virtual courses that have well-defined multistep paths down which the students must follow in order to complete a specific action. A classic example is the path a user must run in order to achieve a given sub-module this is referred to as the learning process from 1st page to the last one, page by page. [14]

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.

The student who easily "gives up" will not fully cover all the pages of the particular syllabus unit (Figs. 4 and 5) 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.










| Page | % of Total Visits That Started Activity | Visits | % of Visit Started | Dropoff % |
|------|---|--------|--------------------|-----------|
| p1 |  | 53 | 100.00 | - |
| p2 |  | 41 | 77.36 | 22.64 |
| p3 |  | 34 | 64.15 | 17.07 |
| p4 |  | 32 | 60.33 | 5.88 |
| p5 |  | 31 | 58.12 | 3.12 |
| p6 |  | 30 | 55.60 | 3.23 |
| p7 |  | 28 | 52.83 | 6.67 |
| p8 |  | 28 | 52.83 | 6.67 |
| p9 |  | 28 | 52.83 | 6.67 |
| ... | | | | |

Figure 4
Processing Sub-module – Digital Image Editing 2










| Page | % of Total Visits That Started Activity | Visits | % of Visit Started | Dropoff % |
|------|---|--------|--------------------|-----------|
| p1 |  | 86 | 100.00 | - |
| p2 |  | 69 | 81.18 | 19.76 |
| p3 |  | 51 | 60.00 | 26.09 |
| p4 |  | 45 | 52.94 | 11.76 |
| p5 |  | 38 | 44.71 | 15.56 |
| p6 |  | 35 | 41.18 | 7.89 |
| p7 |  | 34 | 40.00 | 2.86 |
| p8 |  | 30 | 35.29 | 11.76 |
| p9 |  | 28 | 32.94 | 6.67 |
| ... | | | | |

Figure 5
Processing Sub-module – Digital Vector Graphics Editing 3

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.

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. Web mining methods aim to discover useful information or knowledge from the web hyperlink structure, page content and usage log. 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).

References

- [1] W. Dick, L. Carey, J. O. Carey: *The Systematic Design of Instruction* (7th Ed.). Allyn & Bacon Publishers, Boston, 2008, p. 432
- [2] D. L. Kirkpatrick: *Evaluating Training Programs*. Berrett-Koehler Publishers, San Francisco, 1994
- [3] R. Kosala, H. Blockeel: *Web Mining Research: A Survey*. SIGKDD Explorations, Vol. 2, No. 1, 2000, pp. 1-15
- [4] J. Srivastava, R. Cooley, M. Deshpande, P. N. Tan: *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. SIGKDD Explorations, Vol. 1, No. 2, 2000, pp. 12-23
- [5] O. Etzioni: *The World Wide Web: Quagmire or Gold Mine?* Communications of ACM, Vol. 39, 1996, pp. 65-68
- [6] A. Cohen, R. Nachmias: *A Quantitative Cost Effectiveness Model for Web-supported Academic Instruction*. The Internet and Higher Education, Vol. 9, 2006, pp. 81-90
- [7] C. Pahl: *Data Mining Technology for the Evaluation of Learning Content Interaction*. International journal of E-Learning, Vol. 3, 2004, pp. 47-55

- [8] O. R. Zaïane: Web Usage Mining for a Better Web-based Learning Environment. In: Proceedings of 4th IASTED International Conference on Advanced Technology for Education, Banff, Canada, 2001
- [9] B. Liu: Web Data Mining. Exploring Hyperlinks, Contents and Usage Data. Springer, Berlin, 2006, p. 532
- [10] SPSS Inc.: Web Mining for Clementine 1.5. User's Guide. NetGenesis, Chicago, 2005, p. 89
- [11] I. J. Rudas, P. Tóth: Examination of Learning Activities by Web Mining Methods. In: Proceedings of 9th WSEAS International Conference on Education and Educational Technology, Iwate, Japan, 2010
- [12] C. Romero, S. Ventura, E. García: Data Mining in Course Management Systems: Moodle Case Study and Tutorial. Computers & Education, Vol. 51, No. 1, 2008, pp. 368-384
- [13] G. Piatetsky-Shapiro: CRISP-DM: A Proposed Global Standard for Data Mining. The On-Line Executive Journal for Data-Intensive Decision Support, Vol. 3, No. 15, 1999
- [14] SPSS Inc.: Web Mining for Clementine 1.5. Application template. NetGenesis, Chicago, 2005, p. 177