Searching for indicators in educational processes through the use of a Learning Management System

L. Colazzo & A. Molinari
University of Trento Department of Computer and Management Sciences, Italy

Abstract

Learning processes carried out using ICT leave behind different types of identifying marks or “tracks” in computer systems in general and particularly in the log files of Web servers, the most commonly used system. Our starting assumption is that the study of these traces is essential in order to verify the efficacy and efficiency of these learning processes. In this paper we present the problems found in gathering the data required to analyse such learning processes, and also transforming this data for a multidimensional analysis. With this in mind we built a Learning Management System (LMS) to carry out the distance learning activities of the Faculty of Economics at the University of Trento, and adapted it to our research interests. We also present the multidimensional cube created by our team using the data collected from the LMS, as well as the techniques used for extracting and analysing the data.

1 Introduction

With this paper we wish to show that to collect data on learning processes is a feasible goal, even as the normal activities of an educational institution are taking place. To do this we will refer to our experience of building an e-learning system that has been used for several years by the Faculties of Economics and of Engineering at the University of Trento. In our opinion, the collection of data from an LMS should be included from the initial designing stage. The objective is to use data mining methods [1] to extract useful information from the vast databases of paths that students, teachers and administrative staff leave on log files while using these systems.
LMSs are becoming increasingly popular in university settings for several reasons; on the whole, such a system makes it possible to support both distance education and traditional educational processes based on face-to-face lessons. An important but understudied consequence of using an LMS regards the possibility of obtaining some measurements of the learning process not coming strictly from quizzes, examinations or tests. The fact remains that educational processes are difficult to measure and quantify, and the only way to do it (in situations not supported by technology) is the use of intermediate and/or final tests. We call this new way of collecting information through the interaction with an LMS “involuntary feedback”, as the users do not give explicit feedback on the learning process. In short, the assessment of the learning process provided through testing is of relatively little value in evaluating the quality of the process itself:

- results are affected by cognitive differences among subjects;
- measures are affected by the subjectivity in the evaluation of the teacher;
- performance during tests is conditioned by the ability of the candidate to cope under stress.

The reasons for this weakness in measuring the learning process are obvious: tests measure individual abilities and say nothing about the performance of the educational institution itself. If we calculate the frequency distribution of the students’ performances, and then combine this data with a test taken before or after the learning process takes place, we obtain a meaningless result: few students have low and high performances, while the majority is distributed around the median point. After the entrance test, the position of the Gaussian distribution shows only if the class is high or low. Its possible shifting at the end of the learning process tells us only if the situation has improved or worsened. These results, however, do not lead to an automatic conclusion about the quality of the learning process carried out by the educational institution. Such a conclusion requires more meaningful information, which is more difficult to gather in the absence of an adequately equipped LMS. In other words, it becomes necessary to trace the behaviour of the actors involved in the learning process (students, teachers, tutors, mentors, staff personnel) during the actual process, in order to obtain indications on how to improve the process, and if possible to react in real-time.

2 Measuring user behaviour in an LMS

Electronic communication leaves traces in different forms which for purposes of our study can be used to infer information about the ongoing learning process. A collateral benefit of these “tracks” is that we can collect the data they provide, and then analyse and interpret it so as to better understand and improve the learning process, all of which can be done with or without the conscious participation of the user. This is possible because of the technology on which learning systems are based, generally a dynamic Web application that takes data from a back-end relational database. This integration suggests that we can use data collected by the LMS as a source of information about ongoing learning.
processes; but in order to obtain these results, an LMS must solve three different problems:

- a technological problem regarding the integration of log files from different applications that are not normally integrated;
- a technological problem regarding the presence of useful data in different applications within the academic institution, from places ranging from the student secretary’s office to the dean’s staff.
- A conceptual problem regarding the methods of tracing the process.

As for the first problem, the integration of a Web server log file with the journal of the DBMS is not so easy as may appear to be. In [2] we recently noted why in our opinion it is necessary to provide the LMS with its own log files, where intercepting users interact with the system. Regarding the second problem, academic faculties have many information systems, more or less recognised and formalised as such. The data necessary to create a pervasive informative base supporting the educational process are scattered in different systems, databases, spreadsheets, proprietary applications etc. To complicate the matter, we usually find multiple platforms, different development frameworks and languages, several DBMSs from different vendors, all of which are often managed by different personnel. The third issue is undoubtedly more complex. We are faced with the problem of which metrics to use to evaluate the learning process. A related issue regards the feasibility of executing some observations, and whether some very important phenomena in the learning process are even measurable. In the end, this problem is still not well defined because in some learning processes raw data can not be collected. Our research in this direction is still at a preliminary stage, and for now we are limited to “counting” those things that can be counted.

The WWW has grown exponentially since Tim Berners Lee came up with his original idea for a standard system that allows people to build distributed hypertexts on the Internet, this especially since business has entered the arena. The business approach to the WWW immediately demonstrated the limits of a “static” view of the Web, that is, one of supplying the same pages to different visitors of the site. The need, then, was to overcome the static nature of the HTML language and equip developers with real programming languages able to build dynamic Web sites. In this scenario, the interesting data for the analysis of user interaction are enclosed in two different log files: the Web log files and the journal of the DB that feeds the dynamic Web application. In order to mine these sources they must be integrated, implying several related problems. A DBMS journal records all the activities performed by the DB engine, with a granularity defined by the DB administrator. Thus there are two different possibilities. In the first case, we can move towards a total integration of the different sources in a centralized database on which data mining activities can be performed. In this centralized DB, data must be integrated and normalized in different formats coming from different applications with different data dictionaries, although there are most likely no foreign keys available. In the second case, the code of the dynamic application must be enriched with primitives and methods that capture all the interesting events generated during the user’s navigation. These
relevant events must be recorded in a specialized journal that will be organised accordingly to perform multidimensional/data mining activities. It is also important to consider the granularity and the quality of the information generated by the DBMS journal. Normally, these tools generate atomic information regarding all the transactions performed on the DB, and are generally used in mission-critical applications. On one hand we have log files from the Web server that are too general and contain very little information about activities and, most important, that do not permit a precise identification of the user. On the other hand, data recorded in the DBMS journal are not correlated with Web log files with any foreign key and are also too technical.

The approach taken for this project is the second, which involved enriching the application code with “observation points” where the program creates specific records of predefined format in a specialized log file, from which a datamart will be created (see fig. 1). This task usually requires a certain number of operations: extraction, transformation, cleaning, adaptation to the model of the receiving warehouse, and finally loading. The design of the specialized log files of the LMS should take into account the need to facilitate these operations. The obvious drawback of this approach is the need to intervene on the source code of the system, but at the same time has the advantage of reducing the work of integration among different sources and applications, and especially permits substantial flexibility in the definition of what is needed. This is exactly where our project stands now: we have written the LMS from scratch so we can modify the source code to fit our research needs.

![Figure 1: Data collection in the prototype.](image)

Applying data mining tools and techniques in learning environments is a new thread in the research of distance learning. Data mining techniques are used mainly to support the analysis of student learning processes and the evaluation of the effectiveness and usability of Web-based courses [3]. There are approaches that analyse browsing log files based on data mining technology. The extracted knowledge includes the associative material clusters and the sequences among the material clusters [4]. Other approaches [5] [6] attempt to classify students in
order to predict their final grade based on features extracted from logged data. Yet another approach aims to understand how students and teachers use educational materials, in order to better group these actors into classes with common attitudes and needs [7]. Our objective is to find patterns based on learning behaviours rather than browsing behaviours. In doing this we can help students improve their use of the resource, thanks to the help provided by information on the performance of other students in the same community. The project includes several steps: the first is to gather information to improve the system and its tools: the consequent second step is to see how this has helped to improve actors’ (teachers, students, administrative staff, external personnel, tutors etc.) performances on the system. Another direction we want to explore regards the classification of problems these actors faced while using the system in order to improve the educational material and, specifically, homework tasks.

Pattern recognition has a wide variety of applications in many different fields. The optimal classifier in every case is highly dependent on the problem domain, and therefore it is not possible to use a single classifier that can give good results in all situations. In practice, one might come across a case where no single classifier can classify with an acceptable level of accuracy. We are at a very early stage of the project, and we are making some choices about the right classifier. Probably it will be better to pool the results of different classifiers to achieve the optimal accuracy. In other works comparable to this study, the classifiers used are Quadratic Bayesian classifier, 1-nearest neighbor (1-NN), k-nearest neighbor (k-NN), Parzen-window, multilayer perceptron (MLP), and Decision Tree. These classifiers are some of the common classifiers used in most practical classification problems, and we probably will use these in our experimentations.

3 Summary of functionalities of the target E-learning system

Since building the LMS four years ago [8], [9], we have constantly extended it in order to include new functionalities by using a spiral model. The aim of this system called “OnLine Courses” (OLC) is mainly to distribute learning objects through Internet. These types of system are also known as Courseware Management Systems [10], [11], [12], [13], [14]. The effectiveness of distance learning systems is widely debated, especially in university contexts [15] [16], starting from the first experiences with CBL systems [17] to the advent of Internet [18] and to the modern integration among different media in real-time during the lesson [19]. Currently the system is undergoing considerable change by incorporating some extensions towards mobile technologies, which will heavily modify its architecture. In these situations, an educational institution such as a faculty acts as an aggregate of real communities, where communication technologies act to extend the possibilities of this communication. On these topics, see [20]. The desired result is that at the end of the process we will see growth in the knowledge of all the participants. The LMS used by the Faculty of Economics at the University of Trento is based on the creation of virtual learning communities that are extensions of real communities. Needless to say, in these
communities the teacher and the tutors have duties and rights completely different from those of the students. For each course activated by the Faculty, the homepage for the respective virtual community is activated [21]. Each teacher accesses a personal homepage where all the virtual communities she subscribed to are presented. Each teacher can configure her homepage in such a way that she can access directly the courses attended at that moment. The staff personnel configure the entire environment according to the decisions of the academic body. Figure 2 presents a fragment of the class diagram on which OLC is based. We can see the relationship between members of the faculty and the virtual communities implemented in the system. The LMS supplies a set of tools to the members of the community, tools that facilitate and support the electronic communication. Teachers have tools to self-organise their work, including:

- the calendar of the course;
- the list of all students enrolled in the course;
- indications on the number of download and upload operations done by members;
- export lists of students;
- many bureaucratic requirements such as the classroom register, the final report on the course etc.

Students and teachers are therefore members of the community but with different rights and duties on the utilization of the communication tools. This management of roles, rights and responsibilities has been implemented in the current version of OLC only at the course level: in the new version, this concept has been extended to the whole concept of community and the respective actions that a user can perform. Communities have a hierarchical structure, meaning that a community can contain smaller communities which, in turn, can contain other communities and so on. All the subordinate communities (courses, specializations, work groups, etc.) belong to the root community level (in our case, the Faculty).

Each community has a person in charge, while staff personnel of the Faculty manage the root community. When a person fills the role of coordinator of a community, then she has the right to define which services should be activated in the community. The coordinator also has the right to activate child communities and to choose the new coordinator for them.
As can be seen from this schematic description, the system is quite rich in data structures and functionalities, where different users perform and interact with the system in different ways on different objects. All these activities are monitored and logged, and these constitute the large data warehouse we can build while our users interact with the system. At the moment we have about 5000 students and 400 teachers involved in using the system, and we have an average of 30,000 accesses per month, with at least 500 logged operations per user, without considering the interaction with newsgroups and chats. Including these two tools as well as the tools of the new version, and in addition considering the presence of more sophisticated models of interaction inside the community, we potentially have a huge volume of data with different mining techniques. At the moment, we are analysing some clustering algorithms in order to classify user behaviour in the system.

4 The analysis dimensions

As explained in section 2, it is best to design the log file of the LMS in such a way that it can feed directly the fact tables and dimension tables into a datacube. The fact table includes events, that is, actions performed by the users logged into the system. Every action has some measures associated:

- the number of bytes exchanged during the event;
- two temporal stamps;
- the identification of the community in which the event happened;
- the type of connection and the user that performed the action;
- the identification of the user which performed the action.

The data collected on the user’s behaviour are organised according to five main dimensions: time, community, type of action, role of the user and type of connection. Each dimensions requires some further description.

Time. This dimension has two aspects: the first is obvious and regards the traditional hierarchy of years, semesters, quarters, months, weeks, days, hours, minutes and seconds. The second aspect is specific for each institution and regards the way the organization divides time into academic years, semesters, bimesters etc. This division not only varies among different universities, but also among faculties and even courses. Furthermore, starting and ending dates for the academic periods vary from year to year, and these peculiarities must be taken into consideration while manipulating and aggregating data.

Communities. A user can act within many communities: some are learning communities, but others are of a different nature, such as for example the faculty board. The differences between two communities regard the services available for each of them: a “non-educational” community will not have learning objects, but will include agendas, resolutions, minutes etc..

Typology of actions. Actions performed in the system vary greatly: some include browsing data managed by the system (calendars, programs etc.), others entail a data transfer to and from the system. There are also many actions performed by teachers, for example creating, modifying, using or assessing educational resources. We decided to associate a weight of zero bytes to
browsing actions, while the actions that imply a file transfer to/from the system record the number of bytes transferred. This choice was made in order not to overload the system. Because the system is an application of dynamic pages, the visualization of data requires a query in the DB, and the number of bytes of the results must be calculated each time, thereby augmenting the overload. In file transfer operations, instead, the number of bytes corresponds to the dimension of the file transferred, and this information is already stored in the database.

**The role of the user.** The users of OLC can fill several roles in different communities. For example, a teacher acts as coordinator of the communities related to her courses, and at the same time can be a guest in the communities of her colleagues. Given that the role performed by a user determines the actions that she can perform inside the community, it is necessary to distinguish the role she occupies while performing a given action.

**Type of connection.** The users of OLC can interact with the system from everywhere, and it is useful to discriminate whether they are accessing from the internal network or from an outside provider. Worth noting is that OLC provides explicitly for external users to connect to the system as guests, after they have supplied the necessary data to identify them. A rigorous debate has taken place among our colleagues on this aspect, with two conflicting visions emerging: on one side, those who believe that a large number of guests is an indicator of the quality of the didactics, while on the opposite side there are those who believe that the knowledge included in the system should be accessible only to enrolled, paying students. The solution adopted is a sort of compromise: guests can access but they must be identified explicitly, and in this way, we can model also the dimension of the guests.

The schema of the Data Cube is organised as a snowflake: in figure 3, we present a simplified version of this schema.

![Figure 3: A simplified schema of the cube.](image)

### 5 Some notes on the publicity of the measures

In an e-learning system based on the concept of virtual community, none of the actors is anonymous; everyone knows each other and moreover the system identifies everyone. When an actor enters the system, she is identified and all her actions are traced. This implies that the measures collected during navigation could potentially be used to judge the behaviour of actors. For example, we could identify the students that more frequently use the system, and conclude...
That those students are the most diligent. Obviously such an approach would destroy the validity of the collected data. In our project, the datamart built from the datawarehouse should be constructed so as to defend these aspects of privacy: it should not be possible to analyse data for unplanned, extraneous aims. Therefore, we will provide a diversified vision of the available measures, depending on the actor. Access to the measures is divided into three different categories: private, semi-public and public aggregations. Private measures are offered individually to main actors in the process: students and teachers. These regard specific aspects of the behaviour of the single user of the system. Their use is substantially for self-evaluation. The semi-public measures are offered to the virtual communities and regard specific aspects of the behaviour of the actors in the community. They contain statistical aggregations on the progress of activities within the community, and their use has a twofold aim. First, it allows all the participants to self-evaluate their progress in the learning process, and secondly it enables teachers and tutors to adjust the real process to the learning activities. Public measures regard the progress of all educational activities of the faculty. These are aggregated measures used by the dean and by the Faculty Board to evaluate the quality of their educational offering. At the moment, we have not yet defined precisely which measures will be considered private, semi-public or public. The debate on this aspect is ongoing and involves all members of the faculty.

6 Conclusion and perspectives

The paper presents a project that aims to collect data from a Web-based LMS, offering researchers unique opportunities to study how the actors of such a system interact with it while performing educational processes. For this purpose, we modified an LMS in order to collect data at specific points during user interaction with the system. The prototype provides a way of collecting specialized log files regarding the activities of users in order to identify what approaches to learning lead to success. Web-based systems routinely collect vast quantities of data on user patterns, and data mining methods can be applied to these databases. This paper describes how we designed and implemented the datamart for collecting this data. We are studying data mining methods for extracting useful information from these large databases using online educational resources and how to present this data to stakeholders of the educational process. Currently we are evaluating a series of pattern classifiers and comparing their performance on an online course dataset.

References


