CHAPTER 13

Analysis of user navigational behavior for e-learning personalization

E. Mor, J. Minguillón & J.M. Carbó
Computer Science and Multimedia Studies,
Universitat Oberta de Catalunya, Spain.

Abstract

Personalization is an important issue in e-learning as it might help to improve both student performance and experience of use. In this chapter we describe a framework for studying the navigational behavior of the users of an e-learning environment integrated in a virtual campus. The students navigate through the web based virtual campus interacting with learning resources which are structured following the SCORM e-learning standard. These learning resources are structured following the concept of itinerary which it is basically a temporal scheduling involving several activities and the use of several learning resources. Itineraries may be structured depending on several personalization issues, ranging from student preferences to instructional designer and teacher teams expertise, including also knowledge extracted from the usage in previous semesters. Our main goal is to analyze such user navigational behavior for extracting information that can be used to validate several aspects related to virtual campus design and usability but also to determine the optimal scheduling for each course depending on user profile. We intend to extend the sequencing capabilities of standard learning management systems to include the concept of recommended itinerary, by combining teachers expertise with learned experience acquired by virtual campus usage analysis.

1 Introduction

Web mining is becoming a useful and common tool for institutions, as more and more data is collected from the users browsing the increasing number of web pages with interesting content. The validity of web mining as a tool for extracting useful

WIT Transactions on State of the Art in Science and Engineering, Vol 2, © 2006 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/1-84564-152-3/13

information in any web-based organization system is described in several papers [1–3]. Three types of data are to be managed in any corporation web site: content, structure and usage data, which is the goal of this study. There are several fields where web usage mining can be used for understanding user behavior and navigation [4]. This expertise about users' behavior can be reintegrated in the web-based system (offering user personalized services, for example) in order to improve both user experience and satisfaction, and hopefully, to strength the customer relationship model [5, 6].

On the other hand, e-learning is one of the most promising and growing issues in today's information society. The growth of the Internet is bringing online learning to people in corporations, institutes of higher education, the government and other sectors. The growing need of continuous education and the inclusion of new multimedia technologies become crucial factors for this expansion. The appearance of Learning Management Systems (LMSs) has been a remarkable event for the success of e-learning environments, because there is no longer the need to design specific software for both content delivery and user management.

Several interesting questions arise from the use of web mining techniques in e-learning virtual environments. The possibility of tracking user behavior in such environments creates new possibilities for both web-based system architects and designers, but also for the pedagogical and instructional designers, which create and organize the learning contents. One of the most interesting possibilities is the personalization of the e-learning process. Personalization, which is a term widely used in other environments [7] such as e-commerce, is one of the most well-known and desirable properties of any web-based system, as it pursues the improvement of user experience and satisfaction. Personalization arises from the knowledge extracted from the navigational behavior of the e-learning virtual environment users, mostly students in this particular scenario. In fact, such scenario is a 'closed' system in the sense that every action performed by the users are related to the learning process, and with a set of previously established goals. Therefore, interesting hypothesis about user behavior, navigational patterns and other issues related to the learning process can be formulated and validated by means of web mining tools.

Personalization is a set of technologies and functionalities used in the design of user experiences. The functionality that is part of the personalization can vary, from simply show the user name on a flat web page, to a complex cataloging of the user navigation and the adequacy of products and services based in complex user models [8]. In online learning, personalization is revealed of great utility and importance. The adequacy and adaptation of the learning process it is very interesting as much at educational level as a level from establishing a one to one relationship with the student, and in this sense, it allows to present and offer high quality services and advantages to an every time more satisfied student. It is interesting to note the difference between the personalization and the individualization of the learning process. It is not the invention of this work to individualize the learning process, but to bring a methodology that allows the system to adapt the formative itineraries to the student needs and behavior. The different aspects related to instructional design are not approached directly in this work but it is necessary to mention that

instructional design issues are of special importance and they are not obviated, in the sense that those issues are regarded as fundamental for the personalization system proposed goals. In fact, personalization is not an exclusive issue for information systems, but also for a wide range of applications (see [7] for several examples in other fields).

This work is part of a project concerned with the design of a standards-based e-learning platform that permits the creation of personalize user training itineraries [9], using reusable learning objects as the basic building blocks of the system as well as arriving at a formal methodological and normative specification for automated and semi-automated processes normal for any virtual environment system in the area of e-learning. As regards personalize training itineraries, the basic idea is to convert any current teaching plan (which is usually a completely linear document, static and isolated), into the skeleton of a dynamic and variable process which involves aspects of instructional design for user centered personalization (i.e. the student) and it is related to all the learning objectives which appear throughout an academic period (materials, resources, activities, teaching calendar, etc.), giving rise to what is called a learning itinerary. In this way, learning objects, structured and labeled using standards (LOM for example), are combined according to pedagogical criteria, the know-how of the teaching team, and the recommendations derived from observational studies realized previously with the users of the virtual classrooms (students principally, but also tutors and teachers), creating different possible formative itineraries. The itineraries form a non-linear graph structure which permits the expression of the whole richness of the learning process (obligatory and optional activities, repetition of activities, etc.). Regarding the specification of processes, we will develop a methodology which permits us to define consistent, unambiguous behaviors in learning systems, as currently the concept of itinerary, which are valid for any educative environment based on e-learning. The utilization of the SCORM standard for the representation of these itineraries assures a correct presentation of those contents so that the student has the liberty to advance at his/her own rhythm, but within a framework which has been defined previously by the teaching team.

The system follows all the user actions with two objectives: first, to adapt the learning process according to the rhythm and the actions of each user, as well as the results obtained; secondly, to collect information which can be analyzed later in order to extract useful information for usability evaluation, designing itineraries and measuring the quality of the personalization system, to detect possible problems and unclear points. The adequate combination of learning objects can be automated or semi-automated due to the existence of basic processes formally defined using an ontology and a standard language of definition of learning objects, in such a way that the system can use the process descriptions in the specification of processes that might be automated (to acquire learning processes, or compose new ones based on the learning objectives defined and the needs of the user, to cite some examples). Finally, the usability criteria and interaction between the user and the personalization system is also an important aspect to take into account when designing the system, in order to obtain a balance between privacy, flexibility and supervision, and assure the quality and accessibility of the proposed system.

2 E-learning environments

The intensive use of Internet possibilities not only for content searching and delivery but also for interface design and implementation has completely changed the visions in the open distance education field. E-learning is one of the most promising and growing issues in the information society nowadays. The growth of the Internet is bringing online education to people in corporations, institutes of higher education, the government and other sectors [10]. The growing need of continuous education and the inclusion of new multimedia technologies become crucial factors for this expansion.

One of the most interesting possibilities in any virtual environment is tracking user navigational behavior for analysis purposes, as it may help to discover unusual facts about the system itself. For example, having a user navigational model [11] may be used to perform an automated usability evaluation, detecting whether the system web interface was properly designed or not, and where users find obstacles and difficulties to reach out their objectives. It can be also used to detect bottleneck problems or web areas not used by most users. In the case of a virtual e-learning environment, several analysis levels can be determined and different research questions can be answered with the aid of web mining tools. Instructional designers, teachers and web designers need powerful tools for visualizing all the information collected in a virtual e-learning environment in order to improve the learning process and, thus, user experience and satisfaction, by means of an adaptive [12] environment: monitoring and interpreting the activities of its users [13], inferring user requirements and preferences, and acting upon the available knowledge on its users.

One of the related problems with the design of an e-learning environment is to obtain information about the users' actions, behavior and navigation and about the system usage. As it is described in [1], it is difficult to monitor what the users really do and what is expected they do in the form of navigation and behavior patterns. E-learning environments are usually designed taking usability into account, but it is not easy to determine whether the users feel comfortable with the environment or not, and whether they are capable of carrying out the tasks and actions related with the learning process or not. The extraction of the real learning and training patterns can be of great utility, among other purposes, to determine the degree of quality of the e-learning environment design and to evaluate the concordance degree among the usability requirements and the navigation behavior of the users. For example, it could be very interesting to find correlation among behavior patterns and the instructional and pedagogic design issues. The extraction of behavior and learning patterns must not be used in anyway with the aim of inspecting users and obtaining data that can be used with other goals than just guaranteeing a satisfactory learning and training process. The privacy aspects related to mining user behavior data and system usage data are described in [14], and some real scenario cases can be found in [15]. In Internet and e-learning environments it is not clear yet which indicators, metrics and usability data are suitable and relevant when designing processes and making decisions to design the end-user experience. There are still many design decisions based on hypotheses that are not contrasted with any type

of objective data or real facts. The approach described in this chapter faces these issues and suggests a methodology that can serve to obtain data and results that they provide knowledge about users, their behavior and the environment usage, beyond the usual and common metrics like the hits or the number of accesses to a page [16]. Analyzing the log files generated and stored by an e-learning environment or platform and taking the information about the user's interactions and activities that the system stores, certain navigation paths and patterns can be obtained [17, 18]. Many times, in order to construct accurate paths, the information in the log files should be complemented with other sources of data as embedded marks placed in several points of the virtual environment.

2.1 The UOC virtual campus

The Universitat Oberta de Catalunya [19] (UOC [20], known as Open University of Catalonia in English) is a completely virtual campus which offers 19 official degrees, several graduate programs and post-graduate studies, and a doctoral degree, with more than 35,000 students and more than 1500 people including instructional designers, teachers, tutors, academic and technical staff. The UOC virtual campus is an integrated e-learning environment which allows users to communicate with other users using a mail system and includes an agenda, a news system, virtual classrooms, a digital library and other related e-learning tools. Although the use of classical text printed books is still massive, there is also a growing use of web based e-books and other online learning resources, so the introduction of new e-learning standards such as SCORM [21] is becoming a necessity for maintaining the constant evolution of the virtual campus.

Figure 1 shows the typical initial page that is loaded once an user logs into the virtual campus. Basically, this page includes a left vertical frame with a dynamic menu for accessing all the parts of the virtual campus, a top horizontal frame with recursive navigational buttons for the general and most common actions and, depending on user profile, a set links with all the information of the subjects he or she is enrolled to, and all the latest news in the virtual campus.

Basically, a student taking part in a course has an environment for communicating with the teachers and the other students, a learning plan and a calendar which includes a basic schedule for the activities needed to follow the course. This calendar is a default learning itinerary which is created by the instructional designers and the teachers of the course. The student has access to several learning resources (documents, exercises, etc.) accordingly to such itinerary. Because of the typology of the institution and the courses, which usually are taken in one academic semester (actually, around 15 weeks), the time unit used for designing itineraries is one week. Therefore, when we talk about navigational patterns we need to combine all the different user sessions along the academic semester in a single long-term session, rather than using the information contained in a single login session. It is remarkable that, although it is not possible to predict user behavior for a single session, the set of possible actions and activities is known, as the student navigates through a closed e-learning environment with a specific goal, i.e. to successfully achieve the learning

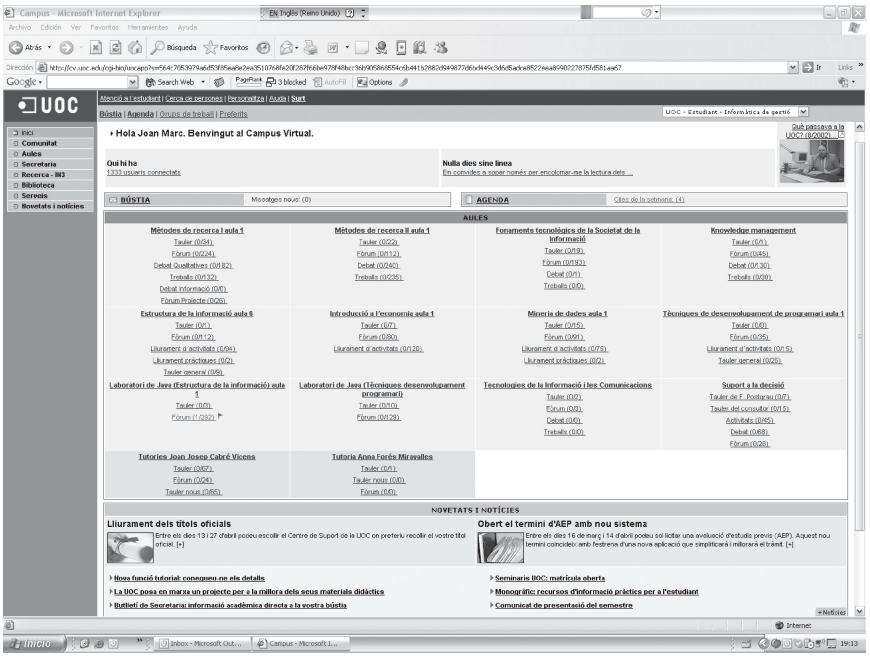

Figure 1: Virtual campus start page once user has logged in.

goals and fulfill the course requirements in order to pass the course exams. This fact can be also used to determine the variables used in the experimental setup. Each course involves a team of instructional designers, a team of teachers, the students which will take part of such course, and a set of learning resources. These learning resources are structured following the concept of itinerary which it is basically a temporal scheduling involving several activities and the use of several learning resources.

The UOC virtual campus is undergoing a structural revision in order to incorporate the use of new e-learning standards for improving user experience by means of personalization [9]. The inclusion of e-learning standards will allow a better tracking of students behavior (using the SCORM 2004 standard [21] capabilities) when accessing the learning resources, shifting also from a blended offline and online learning style towards a more online oriented learning.

2.2 Virtual campus architecture and services

The UOC virtual campus is built upon a complex database server system which uses a hierarchical structure of servers which deal with different kinds of user requests. There are up to 24 front-end servers (depending on the server load at each moment), and an automated load-balance system moves each user login to the front-end with lower load at that moment, switching on and off the total number

of front-ends depending on system load. Other database servers for the digital library, the corporate intranet services, and other management requirements are also connected to the main database server system.

Briefly, the virtual campus uses client-server web technology and common interface to integrate a series of services and functionalities. These functionalities include: access to online educational materials, library resources, and general academic and cultural information; student enquiries management service; and interaction with professors and other students through predefined communication channels (e.g. forums, virtual laboratories, activity spaces). Among others, the following services are offered to students: an email account; a collection of virtual classrooms, where each one has several communication spaces where students and teachers can interact and share learning resources; a digital library which integrates all the digital and non-digital contents into the virtual campus. When users navigate through these services, they leave a track which can be posteriorly analyzed for user modeling purposes. Most of this information is collected by the web servers in the form of server log files, according to the Apache common log format.

3 Navigational behavior analysis

Navigation behavior analysis and user studies may cover a very wide range of potential research and methods, varying from the study of user navigation and choices in a virtual library or in a online store, to the qualitative discovery of user needs and expectations when navigating through a web site. There are many approaches to study and obtain information about user behavior in a computer system or in a web environment. Some of them are quantitative, taking into account the data and events generated and collected by the system [22] these quantitative methods usually focus on statistical analysis using, in some cases, visualization techniques to better understand what users do [23]. Gathering data related with usability parameters and measures may be useful for automating usability evaluation [24] and quality assessment. Other approaches are more qualitative, using methods to handle the complexity of human behavior. In all those different approaches similar issues are addressed and a set of common goals related with usability and with usability data are pursued. Moreover, the combination and fusion of such approaches and methods has been shown its great potential for carrying out user behavior studies [25].

This work focuses the users' knowledge and modeling on the information of their navigation paths and once obtained, the construction of navigation behavior patterns, with the aims of better re-designing the system not only to remove obstacles to the users activities, but also to tailor such design to user characteristics [22]. Following this approach, three different user navigation and behavior patterns levels are distinguished: session level, academic course level and lifelong learning level. Each of these levels of navigation, use and behavior will provide relevant information for constructing the user model and will allow achieving different kind of goals. These three levels arise in a natural way from the use that the students make of the virtual campus and the distinctive dynamics of their learning and training when carrying out online distance studies.

3.1 Navigational levels

Within the virtual campus framework, student behavior may be different depending on the level of analysis that is to be performed. One of the hypothesis that are interesting from a pedagogical point of view is to establish the connection pattern of each student, and to prove that different students follow different connection patterns but that these patterns are limited to a few, mostly because of course structure and temporal restrictions, but also depending on user particularities. In the context of a university, where each subject (several subjects within a course) is taken during an academic semester (i.e. around 15 weeks), two semesters each year, three different navigational levels can be identified: the session level, the course level, and the lifelong level.

The first level, namely the session level, captures the way users navigate with particular goals in mind. For example, how users use the e-mail service or how they access the proposed exercises. At this level, the short-term navigation behavior is studied, i.e. what each individual user does every time that he or she connects to the virtual campus. In this case, a web mining analysis could be helpful to detect problems with the web interface, for automatic usability evaluation purposes [24]. The information obtained at this level will allow validating these and other work hypotheses used for designing formative actions and learning plans. Many times, when designing learning plans, the starting points are hypothesis such as that students connect to the virtual campus in sessions of 20 min, and then they check their personal mailbox first and afterwards they access the courses they have registered for. The usability issues addressed at this level are related to evaluate task flows, detect navigation obstacles, analyze information flows and detect user's most preferred actions and spaces. For example, in the UOC virtual campus, the information provided at this level analysis would display whether students have difficulties when borrowing a book from the digital library, whether they read the academic news and whether the community and social services fit their needs or not.

The second level, namely the course level, tries to join all the single user sessions in a continuous flow during a longer period of time, with a limit of an academic semester. All the aggregated session information will provide a sort of user general course behavior. At this level it is also required to make a follow-up of the different navigation sessions of each user, to observe if each user has similar navigation patterns during a period of time, for example. This medium-term navigation behavior will be useful to validate hypothesis about the relationships of user actions and his or her results, which are related to the way learning resources are organized. The main goals of this level are to determine the navigational patterns followed by users but at a higher scale than in the previous level. For example, it can be interesting to study whether students connect every day or not, or whether they make extensive use of the virtual classroom forums during the weekends or not. All the information collected at this level could be used to feed an intelligent tutoring or adaptive hypermedia system [26]; with personalization purposes.

The third level, namely the lifelong learning level, can be considered a long-term navigational behavior analysis. In this case, the main interest is to analyze

how students evolve from the beginning of a degree until they successfully finish it (or less successfully, they give up). This includes the study of several stages in the student life-cycle: approach and university access, first and following registrations, and so. Performing a data mining analysis at this level could help tutors and mentors to choose more carefully the subjects each student is enrolled to each semester [27]. At this level it may be interesting to discover inappropriate combinations of subjects that might lead to an excessive teaching burden.

In fact, the virtual campus is a rich scenario for experiment design, as different research questions involving different analysis levels can be imagined. Depending on the available information (collected usage data, surveys, academic results, etc.) and the desired goals, different experiments can be designed.

4 Experimental results

In order to test the validity of the assumptions about the navigational behavior of the students in a virtual campus and the connections with their academic performance, an experiment in the course level has been planned. This experiment implies, at it will be shown, to measure several user actions more related to the session level, such as accessing the virtual classroom or the time between consecutive sessions. Two different subsets of 569 and 111 students taking a degree in Computer Science, who enrolled in the subjects 'Foundations of Programming' (an introductory course to programming for new students in their first semester) and 'Compilers I' (a course for advanced students), respectively, have been selected as the matter of study. These students may also be enrolled in other subjects, but we will focus in the navigational actions related to the subject matter of study. Nevertheless, for a real personalization scenario all user actions should be taken into account for the profile analysis, but with an attempt to find a trade-off between model accuracy and complexity.

Basically, students connect to the virtual campus and access the virtual classroom to follow the learning activities designed for each subject, according to a previously established scheduling. In the case of the subjects studied in this chapter, students are asked to solve an optional exercise which is published during the first week (once the course has started) and that it must be solved and returned back to the teacher within 12 days (including one or two weekends, depending on each course). Students have specific places for both accessing the exercise description and rendering their solution. These spaces can be identified in the log file, so the exact moment when students perform the action of exercise download or upload is known. It is worth mentioning that this exercise is not mandatory, but it is strongly recommended as the final subject evaluation can be broken apart in several continuous evaluation activities such as the proposed exercise. Therefore, all students are supposed to follow the proposed activities, because those who do not follow them must take a final exam at the end of the semester, which usually has a higher degree of difficulty. We are interested in studying which students decide not to do the first exercise (or if they do it, but with poor results) in order to see whether such failure is somehow related to the way they navigate through the virtual campus and to their socio-demographic background. As mentioned before, this information can be collected

Table 1: Marks obtained by the students in the first exercise.

Mark	A	B	C+	C–	D	N	Total
Foundations of Programming	216	26	63	0	49	215	569
Compilers I	22	65	12	1	0	11	111

online by an intelligent tutoring system and helps each student fulfill his or her learning path much better, under a improved and personalized learning process.

Table 1 shows the results obtained by the students in the first exercise (it is worth mentioning that it is a simple exercise to introduce them to the subject of study, with a medium degree of difficulty, so it is not expected that many students will get poor marks, i.e. 'C−' or 'D', but on the contrary more students are not doing it, i.e. 'N'). Notice that, as expected, figures are different depending on the subject. For 'Foundations of Programming', 216 out of 569 students (37.8%) do not make the proposed exercise, and 49 more present a very poor exercise (8.6%). This is a well known fact for this course, so any information about the profile of users failing in this exercise would be very helpful. On the other hand, for 'Compilers I', only 11 students decided not to do the proposed exercise, and only one did it poorly and fails.

4.1 Server log files

For each action a user performs in the virtual campus, one or more lines representing such actions are logged in several servers. Furthermore, depending on the type of action, several servers might log the same action but using different information. In this work we have used mainly the log files from the Apache servers which act as front-ends, once they have been joined in a single file, generated every day. This file is firstly pre-processed in order to remove all those log lines which are surely not hits produced by the user, such as the load of icons, style sheets, banners, and so. This pre-processing reduces the amount of lines in a 90%. Nevertheless, during a typical week, the total number of lines that needs to be processed is still about 24 million, approximately 12 GB, which is a very large figure. Therefore, a second pre-processing, more oriented towards narrowing the experiment, is required, as described in [28].

Users can be uniquely identified because there is a unique session number generated each time a user logs into the virtual campus using his or her username and password. IP addresses are discarded because there is the possibility of many user accessing through the same proxy server which might mask the real IP address. Therefore, it is possible not only to identify individual users but also each individual session, which is useful to establish the different navigational levels described in the following section. When the user browses areas of the virtual campus where the session number is not required (public areas, for example), it cannot be successfully tracked, so those lines without session number are also removed.

All user interactions with the virtual campus are logged by one or more servers (web servers, database servers, etc). As we have stated before, the UOC's virtual

campus has 24 Apache front-end web servers. Our work is currently focused in analyzing the standard Apache server log files generated by these 24 front-end servers. The variables recorded in the log files are: originating request IP number, local date and time, request URL and referrer URL. To facilitate tracking user and session information all the virtual campus features that require user authentication use an encrypted string embedded into the request or referrer URLs. These are the only interactions we take into account in our analysis because the remaining requests can not be traced down to a user or session context.

To deal with privacy issues, the encrypted string carrying user and session information is removed from the original requested and referrer URLs and substituted with a new user and session identifiers that cannot be traced back to the original user. These new user and session identifiers have a one-to-one map with the original users and sessions. This mangling process enhances privacy while retaining the ability to cross join user interaction data with other user demographic and academic data obtained from UOC administrative databases. Although this study may raise several concerns about privacy issues [14], all collected information is used only for academic purposes. All students are aware of the fact that all their interactions within the virtual campus are being logged and that all private record data is used only within the institution.

The analysis process has several steps, as outlined in Fig. 2. This process is only partially automated. Each day, all the log files that come out of the 24 front-end servers are joined into a single log file. Next, the request lines that have no useful information from the user interaction point of view – image, cascading style sheets, etc. – are removed from the original log file. All lines that lack user and session information (as we have pointed out above) are also removed. This step reduces the total size of the log file by a 90%. After this initial reduction, the average size of the log file is still 12 GB or 24 million lines per week.

Depending on the kind of analysis we want to undertake, this stripped down log file is further reduced by selecting the interaction lines on a user basis. In order to perform the experiments described in this chapter, we have used two different student sets. The first set stores all the interactions of a group of 569 students enrolled in a course named 'Foundations of Programming', with dates ranging from

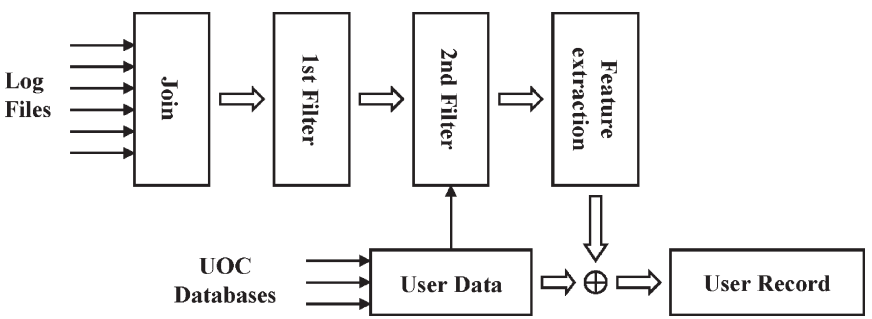

Figure 2: Preprocessing steps for obtaining user records.

23 February 2005 to 11 March 2005. The filtered log file for this set is 1,300,000 lines long. The second set stores the same information a group of 111 students enrolled in a course named 'Compilers I', with dates ranging from the same initial date to 17 March 2005. In this case, the filtered log file for this second set has around 220000 lines. The intersection between the two sets is void, as they are students from different degrees.

The rationale for selecting these sets of students and the related timeframes is the following. First, the datasets are small enough to allow us to experiment with different analysis methodologies. Second, the time frame starts at the very beginning of the course (23 February) and goes on till the students are requested to deliver their first practical exercise (deadline 12 or 17 March, depending on the course). The practical exercise (named PAC in UOC terminology) is of moderate difficulty, as it server as an introductory exercise according to the UOC pedagogical model [19]. As mentioned before, the submission of the PAC is optional but scoring a high mark helps the students in their final course performance evaluation, whilst not doing or failing the PAC has no negative consequences in the final course mark. Finally, this data comes from a closed e-learning environment where all user actions relate to the e-learning goals. Other user data like number of subjects enrolled during the term, number of book loans requested, total number of terms attending UOC is also readily available.

4.2 Data pre-processing and feature extraction

The experiments are performed using a reduced log file which removes all the useless information present in the Apache common log file, filtering out also those students not enrolled in each course matter of the study. The final log files have 13,000,000 and 220,000 lines approximately, which are reasonable data set sizes for studying user navigational behavior in a focused environment. Figure 2 shows all the pre-processing steps, starting from the original log files and databases, until the final data set is generated.

The main advantage of being in such a closed environment (the virtual campus) is that there are other available user data which may be relevant for analysis purposes. For example, from the transcript of each student we can extract the total number of subjects he or she is enrolled in, the number of semesters he or she has been studying in the UOC, and so on. Other information, such as the number of book loans requested or user satisfaction surveys are also available. Although this study may give rise to several questions about privacy issues [14], all the information collected about students is used only for academic purposes. Furthermore, all the students are aware of the fact that all actions in the virtual campus are logged, and that all private record data is used only within the institution. After the initial pre-processing, two filters are applied one for each of the session and course levels of analysis.

The first filter applied to the raw log file is the session filter. The aim of this first filter is to aggregate all user requests within a session into a single set of variables. The initial problem we face when we try to perform a session level analysis is clearly identifying the start and the end of a particular session. We have identified three

kinds of sessions according to their start-end pattern: regular, aborted and fuzzy. Regular sessions are those sessions where the user starts interacting with the virtual campus and he or she is assigned a new session identifier. When the user decides to end the session he or she presses the virtual campus exit link. Aborted sessions are like regular sessions but do not end with the virtual campus exit link, the user ends just by closing the web browser. Aborted sessions need to be summarized at the end of the log file sequential analysis. Fuzzy sessions are, in fact, artifacts from the current virtual campus setup. Due to the load balancing nature of the web server layer, and its multi-threaded behavior, some users are not assigned a new session number each time they login; instead, they retain their previous session identifier. Thus, the only way to track session start and session end in this context is to account for inactivity intervals. When the lap between two consecutive session lines is greater than 20 min, a new session is accounted for regardless of the session identifier.

The first filter algorithm sequentially walks through the log file recording the following primary and derived session variables: user identifier, session identifier, session start local date and time, session end local date and time, user's hit count, session duration in minutes, day of week at the start of the session and hour at the start of the session. These are what we call generic variables. We also take into account some extra variables that are only relevant to a particular experiment at hand. The analysis that we present in this chapter takes into account several additional variables: the number of messages posted to the course forum and virtual laboratory during the session, local date and time of exercise proposal download if any, and local date and time of exercise resolution submission if any. In the next section we will explain why these extra variables are relevant in the context we are presenting. Identifying these extra variables usually implies writing specialized code modules to undercover the events to be taken into account. We use the extensive regular expression facilities of the Perl language to help extracting this kind of information out of the raw URLs. The asymptotic time cost of this first stage algorithm is linear to the amount of log lines and the asymptotic space cost depends almost exclusively of the amount of aborted sessions because they need to be processed at the end of the sequential walk.

The user level analysis follows the session level analysis (see Fig. 2). The aim of this second filter stage is to collapse all session tuples into one single course tuple for each user. Therefore, we will use the following variables as the input for the classification and clustering algorithms:

- GENDER: Although it is usual to have much more men studying engineering degrees than women, it is interesting to include it in this study to confirm the intuitive idea that gender is unimportant.
- AGE: This variable might provide important information about the socio-demographical background of students. For example, older students usually have greater family obligations than younger ones.
- NEW: Whether the student is new to the UOC virtual campus or not. New students may experience difficulties using the virtual campus, so it is interesting to validate such hypothesis.

- FIRST: Whether the student takes the course for first time or not. The students that failed to pass the course in the past semester are more likely to behave differently because they have information about their own experience.
- TOTALCREDS/ADAPTEDCREDS: The total number of course credits which the student is enrolled to / adapts from previous studies. The TOTALCREDS variable is directly related to the amount of time that the student needs to dedicate in order to follow all the proposed activities, so students with a large amount of credits are more likely to drop out from one or more courses, thus not doing the proposed exercises.

This information will be combined with the navigational behavior extracted from a very basic analysis of their navigational patterns during the period of time determined by the course starting and the day after the first proposed exercise must be rendered:

- *A set of information related to the session level*: the number of total sessions in the virtual campus (TOTALSESS), the mean delay between two consecutive sessions (MEANINTDUR), the mean length of each session (MEANDUR), and the mean number of hits (user-driven actions) in each session (MEANHITS). Although the exact intention of user actions in each session is unknown, these variables describe a basic navigational pattern in the period of time which is being analyzed. In order to study student habits, the number of total of sessions is also computed for each day of the week, creating seven new variables WD_i ($i = 1$ for Monday and $i = 7$ for Sunday). Then, a simple index measuring whether the student connects preferably on weekends or not is computed as $(\mathrm{WD}_6 + \mathrm{WD}_7) \big/ \sum_{i=1}^{i=7} \mathrm{WD}_i$ (namely WEEKENDPCT).
- *A set of information related to the course level*: the number of messages posted in the appropriate virtual classroom forum (FORUM), the number of messages posted in the associated laboratory forum (LAB), and the delay between the moment that the proposed exercise is published and the moment student accesses to its content (DELAY).

Therefore, a total of 21 variables are used for clustering and classification purposes. We are interested in somehow predicting which mark will have a given student, or at least, whether he or she will pass or fail the proposed exercise. For the course 'Compilers I', as only one student presents a poor exercise (marked with a 'C−'), this is almost equivalent to study whether a student renders or not the proposed exercise. On the other hand, for the course 'Foundations of Programming' there is an important set of students which are marked with a 'D', showing a very different behavior with respect the other set. It is worth mentioning that the global aim of this study is to understand user navigational behavior and to explain some well-known facts beyond intuition, but no building an accurate system for predicting a particular scenario such the described experiment. Therefore, any personalization effort must take into account not only students behavior but also the whole learning context.

It is worth to mention that there is navigational data for all the students in the 'Compilers I' course, but this is not the case for 'Foundations of Programming', as 25 students (4.4%) never connect to the virtual campus, and obviously they do not do the proposed activities. Therefore, the experiments described in the following section are performed using only the data of the other 544 students. During the pre-processing stage, it was detected that one of this 25 students had a 'A' mark for the proposed exercise, showing that the teacher made a mistake during the process of introducing the qualifications into the system. This illustrates the necessity of a higher degree of control by the system itself in order to avoid human mistakes, i.e. if a student has not uploaded the proposed exercise, automatically mark him or her with an 'N' without any teacher intervention. This case is a good example of how usability, utility and personalization may be improved by analyzing the data obtained from the virtual campus users activity.

4.3 Web mining

Once the data described in the previous section has been tabulated, and a single record describes the collected data for every student, several data mining techniques can be applied. Among others, unsupervised clustering by means of the TwoStep algorithm [29] and supervised classification by means of classification and regression trees [30] are the most useful because of the interpretability of the obtained results, despite the fact that both techniques might not achieve the optimal classification accuracy.

4.3.1 Variable relevance

Decision trees can also be used to measure variable importance, as suggested by [30]. Although several methods have been proposed in the literature [31], we will use one developed by the authors which tries to exploit decision tree diversity when trees are built from similar sets created using a bagging approach [32]. Basically, this method builds a large number of similar decision trees, one for each possible bagging training set, and then variable relevance is computed by weighting the number of times each possible classification feature is selected and its position in the decision tree, giving more importance to variables near the root of the tree.

For the 'Foundations of Programming' course, the most important variables are DELAY, TOTALSESS, and surprisingly, WD_5 (the exercise must be delivered on this day, though) and WD_2. The variable MEANDUR also deserves a special mention. Regarding the 'Compilers I' course, the most important variables are MEANINTDUR, TOTALCREDS, TOTALSESS, MEANDUR and DELAY. In this case, as most students do the proposed exercise, DELAY does not become so relevant as in the other course. It is also surprising that interaction variables (FORUM and LAB) are not considered important for classification purposes. On the other hand, GENDER and ADAPTEDCREDS are the least relevant variables, as expected. Both WD_6 and WD_7 are also considered irrelevant, which is also surprising, as most students are supposed to have little time for studying during the

working week. Therefore, the hypotheses about user interaction during the weekends must be revised.

This simple experiment shows that the same classification features are not relevant for different data sets, although some variables (TOTALSESS, MEANDUR, DELAY) seem to be more robust for describing user behavior even for different learning contexts. DELAY is obviously a variable which contains relevant information which may be used for personalizing the course level: if a student waits too long to download the proposed exercise, he or she will be probably not able to finish it or to obtain a good mark. Therefore, an automatic system response could be designed to warn students (or their teacher) that they are approaching a deadline (a threshold, for example). This threshold could be estimated by combining teacher expertise and previous results. For example, a personalized banner or a tailored mail could be helpful to make the student aware of the proposed exercise and, optionally, downloading it or discarding it.

4.3.2 Unsupervised clustering

The second experiment tries to group students according to their navigational behavior, without taking into account student's socio-demographic background or the mark that they obtain. The TwoStep algorithm is used to discover patterns in the set of input fields. Records are grouped so that records within a group or cluster tend to be similar to each other, but records in different groups are dissimilar. The number of clusters is automatically selected. This study tries to identify which variables are relevant for classification purposes but from a different approach. A posterior supervised classification analysis could be then devised to design a recommendation system or an adaptive system for tutoring purposes combining both approaches.

For the 'Compilers I' students, two clusters are generated, with 27 and 84 records, respectively. The most relevant variables for clustering purposes are TOTALSESS, MEANINTDUR (which are obviously correlated), FORUM and LAB, which are significant at $p < 0.001$, and DELAY, which is significant at $p < 0.05$. MEANHITS, MEANDUR (both are strongly correlated) and WEEKENDPCT are not significant at all. These two clusters capture very well students' interactions: the students who connect more irregularly to the virtual campus do not post messages in the classroom spaces, and DELAY is also higher for these students than for the rest. On the other hand, the 'Foundations of Programming' students are grouped in three clusters, with 91, 318 and 135 records, respectively. In this case, user behavior is so different that all variables become significant at $p < 0.001$ except WEEKENDPCT, which is not relevant at all. Once again, though, students with a higher interaction pattern obtain better results than the rest (cluster 1 in Table 2), while the other two clusters show different values for DELAY, for example.

Table 2 shows the marks obtained by the students in each cluster. Notice that for 'Compilers I', the students who render a simple solution (a 'C+' mark) have a navigational behavior more similar to those who decide not to do it than to those who successfully solve the exercise. For 'Foundations of Programming', this separation is not so clear and needs to be analyzed more deeply.

Table 2: Marks distribution according to the obtained clustering.

	Foundations of Programming						Compilers I					
Cluster	A	B	C+	C–	D	N	A	B	C+	C–	D	N
1	69	4	9	0	4	5	11	13	2	0	0	0
2	127	16	44	0	38	93	11	52	10	1	0	11
3	19	6	10	0	7	93			–			

4.4 Data fusion

At present, UOC is introducing the use of the SCORM 2004 standard [21] for both course development and tracking purposes, as described in [9]. Nevertheless, other LMSs also incorporate the tracking capability, widening the options for obtaining data about the students, such as WebCT [33, 34]. Furthermore, when students browse other virtual campus services such as the digital library, for example, they also leave a track that can be used for personalization purposes [35]. Therefore, a reasonable data standardization process is needed to ensure that all these data can be properly described [36]. This process, which is called 'data fusion' based on the nomenclature used in other fields (e.g. remote sensing), might be very useful for overcoming all the classical log files problems: user identification, huge size, lack of user goals, etc.

In fact, the combination of different navigational strategies for the same goal (distance learning using a virtual campus with blended online and offline activities) will change the way students interact with the virtual campus and, therefore, the learning context. Moreover, having SCORM compliant courses will generate enough usage and academic data, which accurately analyzed will serve to adapt and personalize the learning process. These new personalized courses may lead to new ways of interaction and learning contexts and, hence, the system usage data collected could be different and consequently new processing and analyzing methods will be required. Simultaneously, an adaptive system for processing, analyzing and personalization purposes will be necessary.

5 Conclusions

In this chapter we have described an analysis performed in the UOC virtual campus with the aim of studying the relationship between user navigational patterns and the academic results achieved by the students enrolled to several subjects in computer science, namely 'Foundations of Programming' and 'Compilers I'. Three possible levels of analysis are described, and an experiment designed for the course level is outlined to show the possibilities that arise from the use of web mining tools in an e-learning environment for personalization purposes. Although the results shown in this chapter are preliminary and they are part of an ongoing project, it

is worth to mention that some intuitive ideas that the teachers and instructional designers have about users navigational behavior can be validated with a simple clustering analysis. Obviously, a deeper analysis is required to better understand the complexity of the actions taken by the students. Results show that even a simple analysis may be useful to determine which variables are relevant for both clustering and classification purposes (for example, all the variables related to the interaction patterns), while other variables are not relevant at all (the percentage of sessions during the weekend, for example).

Further research in this area should include the use of other clustering and classification techniques for extracting information relevant to the learning process. The inclusion of other variables which may be also relevant to study user behavior may also improve both prediction accuracy and results interpretation. The extension of this study to other subjects with larger subsets of students or with different background (taking a degree in Social Sciences, for example) is also under consideration, specially for subjects with a known poor academic performance, as such criterion is directly related to user satisfaction. Finally, data fusion from different sources (web logs, internal marks, external databases, e-learning standards tracking tools) is also an interesting possibility.

Acknowledgments

This work is partially supported by the Spanish MCYT and the FEDER funds under grant TIC2003-08604-C04-04 MULTIMARK.

References

[1] Srivastava, J., Cooley, R., Deshpande, M. & Tan, P.N., Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations*, **1(2)**, pp. 12–23, 2000.

[2] Kosala, R. & Blockeel, H., Web mining research: a survey. *SIGKDD Explorations*, **2**, pp. 1–15, 2000.

[3] Zhang, F. & Chang, H.Y., Research and development in web usage mining system-key issues and proposed solutions: a survey. *Proc. of the 2002 Int. Conf. on Machine Learning and Cybernetics*, Vol. 2, pp. 986–990, 2002.

[4] Chi, E.H., Pirolli, P. & Pitkow, J., The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, ACM Press: New York, NY, pp. 161–168, 2000.

[5] Spiliopoulou, M., Web usage mining for web site evaluation. *Communications of the ACM*, **43(8)**, pp. 127–134, 2000.

[6] Marsico, M.D. & Levialdi, S., Evaluating web sites: exploiting user's expectations. *International Journal of Human-Computer Studies*, **60(3)**, pp. 381–416, 2004.

[7] Riecken, D., Personalized views of personalization. *Communications of the ACM*, **43(8)**, pp. 27–28, 2000.
[8] Kramer, J., Noronha, S. & Vergo, J., A user-centered design approach to personalization. *Communications of the ACM*, **43(8)**, pp. 44–48, 2000.
[9] Mor, E. & Minguillón, J., E-learning personalization based on itineraries and long-term navigational behavior. *Proc. of the Thirteenth World Wide Web Conference*, New York City, NY, Vol. 2, pp. 264–265, 2004.
[10] Rosenberg, M.J., *E-Learning: Strategies for Delivering Knowledge in the Digital Age*, McGraw-Hill, Inc.: New York, NY, 2002.
[11] Kay, J. & Lum, A., Creating user models from web logs. *Proc. of the Intelligent User Interfaces Workshop: Behavior-Based User Interface Customization*, 2004.
[12] Paramythis, A. & Loidl-Reisinger, S., Adaptive learning environments and e-learning standards. *Electronic Journal of e-Learning*, **2(2)**, pp. 181–194, 2004.
[13] Thomas, R., Kennedy, G., Draper, S., Mancy, R., Crease, M., Evans, H. & Gray, P., Generic usage monitoring of programming students. *Proc. of the ASCILITE 2003 Conference*, Adelaide, Australia, pp. 715–719, 2003.
[14] Clifton, C. & Estivill-Castro, V., (eds.), *Proc. of the ICDM 2002, Workshop on Privacy, Security and Data Mining*, Vol. 14, ACS, Maebashi City, Japan, 2002.
[15] Spinello, R.A., *Case Studies in Information Technology Ethics*, Prentice Hall: Upper Saddle River, NJ, 1996.
[16] Spiliopoulou, M., Pohle, C. & Faulstich, L., Improving the effectiveness of a web site with web usage mining. *Proc. of the Int. Workshop on Web Usage Analysis and User Profiling*, Springer-Verlag: London, Vol. 1836, pp. 142–162, 1999.
[17] Bucklin, R.E. & Sismeiro, C., A model of web site browsing behavior estimated on clickstream data. *Marketing Research*, **XL**, pp. 249–267, 2003.
[18] Cadez, I.V., Heckerman, D., Meek, C., Smyth, P. & White, S., Visualization of navigation patterns on a web site using model-based clustering. *Proc. of the 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Boston, MA, pp. 280–284, 2000.
[19] Sangrà, A., A new learning model for the information and knowledge society: The case of the UOC. *International Review of Research in Open and Distance Learning*, **2(2)**, pp. 1–19, 2002.
[20] UOC, http://www.uoc.edu
[21] ADL, Sharable Content Object Reference Model (SCORM) 2004, 2nd edn, overview, 2004.
[22] Hilbert, D.M. & Redmiles, D.F., Extracting usability information from user interface events. *ACM Computing Surveys*, **32(4)**, pp. 384–421, 2000.
[23] Chi, E.H., Improving web usability through visualization. *IEEE Internet Computing*, **6(2)**, pp. 64–71, 2002.
[24] Ivory, M. & Hearst, M., The state of the art in automated usability evaluation of user interfaces. *ACM Computing Surveys*, **33(4)**, pp. 173–197, 2001.

[25] Juvina, I., Trausan-Matu, S., Iosif, G., Veer, G.v.d., Marhan, A. & Chisalita, C., Analysis of web browsing behavior – a great potential for psychological research. *Proc. of 1st Int. Workshop on Task models and Diagrams for user interface design*, Bucharest, Romania, 2002.

[26] Brusilovsky, P., Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, **11(1–2)**, pp. 87–110, 2001.

[27] Tattersall, C., van den Berg, B., van Es, R., Janssen, J., Manderveld, J. & Koper, R., Swarm-based adaptation: Wayfinding support for lifelong learners. *Proc. of the Third Int. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems*, Eindhoven, The Netherlands, Lecture Notes in Computer Science, Vol. 3137, pp. 336–339, 2004.

[28] Cooley, R., Mobasher, B. & Srivastava, J., Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, **1(1)**, pp. 5–32, 1999.

[29] Zhang, T., Ramakrishnan, R. & Livny, M., BIRCH: an efficient data clustering method for very large databases. *Proc. of ACM SIGMOD Conf. on Management of Data*, Montreal, Canada, pp. 103–114, 1996.

[30] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., *Classification and Regression Trees*. Wadsworth International Group, 1984.

[31] Duda, R.O., Hart, P.E. & Stork, D.G., *Pattern Classification*, 2nd edn, John Wiley & Sons: New York, 2000.

[32] Breiman, L., Bagging predictors. *Machine Learning*, **24(2)**, pp. 123–140, 1996.

[33] Marquardt, C., Becker, K. & Ruiz, D., A pre-processing tool for web usage mining in the distance education domain. *Proc. of the Int. Database Engineering and Applications Symposium*, pp. 78–87, 2004.

[34] Mazza, R. & Dimitrova, V., Visualising student tracking data to support instructors in web-based distance education. *Proc. of the 13th Int. World Wide Web Conf.* (alternate track papers & posters), ACM Press: New York, NY, USA, pp. 154–161, 2004.

[35] Ferran, N., Mor, E. & Minguillón, J., Towards personalization in digital libraries through ontologies. *Library Management Journal*, **26(4/5)**, pp. 206–217, 2005.

[36] Avgeriou, P., Papasalouros, A., Retalis, S. & Skordalakis, M., Towards a pattern language for learning management systems. *Educational Technology and Society*, **6(2)**, pp. 11–24, 2003.