Discovering Student web Usage Profiles Using Markov Chains

Alice Marques and Orlando Belo
University of Minho, Portugal
alicemarques5@gmail.com
obelo@di.uminho.pt

Abstract: Nowadays, Web based platforms are quite common in any university, supporting a very diversified set of applications and services. Ranging from personal management to student evaluation processes, Web based platforms are doing a great job providing a very flexible way of working, promote student enrolment, and making access to academic information simple and in an universal way. Students can do their regular tasks anywhere, anytime. Sooner or latter, it was expected that organizations, and universities in particular, begin to think and act towards better educational platforms, more user-friendly and effective, where students find easily what they search about a specific topic or subject. Profiling is one of the several techniques that we can use to discover what students use to do, by establishing their user navigation patterns on Web based platforms, and knowing better how they explore and search the sites' pages that they visit. With these profiles Web based platforms administrators can personalize sites according with the preferences and behaviour of the students, promoting easy navigation functionalities and better abilities to response to their needs. In this article we will present the application of Markov chains in the establishment of such profiles for a target eLearning oriented Web site, presenting the system we implemented and its functionalities to do that, as well describing the entire process of discovering student profiles on an eLearning Web based platform.

Keywords: web based elearning platforms, web usage profiling, Clickstream analysis, Markov chains, Navigation paths analysis

1. Introduction

The development of a web site by an organization may not be enough to obtain the success it expect on the Internet. Being the web a market with a large facility to publish contents, certainly will exists other sites, with the same or very similar contents, that users can use as an alternative. Thus, it's necessary to ensure that the site is in line having what its potential users want, in order to avoid they abandon it after some time of using, going to another site with better contents proposals, a more adjusted organization to their needs, or incorporating strong user friendly navigation facilities.

During the last few years, the number of students using web based platforms for learning activities grew a lot, reaching a level quite interesting to analyse the way how and when they use those platforms and their resources. Goals are very clear: to optimize quality of service, to increase resources availability, to attenuate the effect of services (and resources) not used, and to facilitate access to didactic material 24 hours a day. Today, we see universities promoting the development and maintenance of very effective sites, being their administrators quite concerned about their use and effectiveness. Some of them use to monitoring the activities performed on their sites, especially the ones that support eLearning platforms, in order to provide better quality of service, to supply better information (on time) and to avoid eventual system's downtimes. Additionally, this way of looking for web usage has been improved in the sense to discover critical periods of site navigation and exploitation, trying to prevent service bottlenecks and lower rates of service quality – these are two of the most critical aspects that lead students to avoid or abandon eLearning oriented Web sites.

For a university site be successful, we believe we must know, as in traditional business areas, the type of users that navigate in the site, once they are the primordial elements that will influence site advancement, as well as the evolution of the organization itself. That is why more and more organizations try analyzing several questions, related with the form how users interact with the site - which pages are more and less visualized, what are the main pages of entry and exit the site, which are the hours with more traffic, the frequency with which a user visits the site, and so on. The analysis of this type of questions will identify student usage profiles that use a particular eLearning oriented site. The work of identifying usage profiles is a systematic activity, involving massive exploitation of all the available clickstreams, storing, transforming and analysing them with the most recent techniques to do that. With these profiles established we could personalize a site according with the needs of students, promoting easy navigation and better ability to response to customers needs.
To establish student profiles, identifying their behaviour when using a eLearning site, we need to monitoring their daily usage and collect all information we can about their activities inside the site,

knowing what kind of links they use to follow, services or documents they use to access, their frequency of usage, or simply what are their usual entry points. All of this can be done observing the access logs of the site, and analysing things such as the IP number of the machine used during the navigation process, the pages visited (their links, and data and time of access), time of permanence in a particular page, and the resources used, just to name a few. Taking as the IP address as the basis to characterize student navigation, we can ordered the pages by date and time of access and group them into sessions, considering a predefined navigation period, which will define the period to group the visited pages, and a period of inactivity that will define the end of a particular session. Then, analysing all the student sessions we can establish navigational patterns, representing the most regular paths that students use to follow during a session, and consequently generate a profile to personalize (if necessary) the contents of a site.

In this paper we will present the application of Markov chains in the establishment of usage profiles for a target eLearning oriented Web site. With the profiles established we can identify the type of user that is visited and with this knowledge we can provide to students in a better way the services that they probably will use in a near future. Additionally, we will present the way we collected the data, designed and explored the mining models, and the way that Markov chains conduced us to the establishment of the student web usage profiles. We finalize the paper with some conclusions and presenting some paths for future work.

2. Using web based eLearning platforms

2.1 Overview and general features

Web based platforms are expanding their influence in many sectors. From services to industrial plants we can find several applications in the area. ELearning (Carliner & Shank 2008), as we know, it is not an exception. For many years educational institutions began to install and explore Web based platforms, frequently only as a simple way to promote them selves or to receive documents repositories. Quickly they discover the huge attractively and flexibility of Web based platforms, a great freedom in courses' contents managements and maintenance, and as a practical way to improve learning processes, share knowledge and expertise, augment student enrolment, and of course reduce operational and maintenance costs.

With the emmergence of Web based eLearning platforms new horizons has been open to the implementation of new studying environments. These platforms help educational institutions training diverse student populations and provide a very attractive way to enlarge classrooms, taking them where students are. Today this is not a novelty it is a clear reality. The improvement of the Internet and its services helped a lot on this process. Its universal access and its easy usage facilitate a lot the adoption of these platforms. It is a market in a continuous expansion, and it is easy to see that for the number of educational tools that appear every single year. Today, the difficulty is not their use and exploitation but its selection and adoption. Universities have now a lot of options concerning web based educational platforms. Model[1], Blackboard[2], Claroline[3] or Atutor[4] are only a few references that we can find in the market. The services they provide are quite diversified. All of them work towards helping educators to create, manage and maintain online courses, and all their related services, for large communities of students, having abilities to cover educational topics from primary schools to universities. To do their jobs well, Web-based eLearning platforms provide a set of features very powerful, which includes today things such as integration of multimedia objects, multilingual support, project management tools, data import and export services, personalised access based on role definitions, activities reports, evaluation tools, or heterogeneous document types hosting, just to name a few (Tucker et al. 2002). There is an evidence that Web-based ELearning systems (Schewe et al. 2005) are a clear reality with a significant impact in our lives. We face them practically in any educational institution supporting current activities and adding new value to personal education and student enrolment (Hosan et al. 2006). Day after day, educational managers give more attention to these platforms stimulating and supporting their design based on student profiles, and creating flexible and friendly navigation structures, as well are creating methods and processes to model new Web based educational systems (Rokou et al. 2004). Profiling is one of the best ways we have to go towards an adaptive Web-based eLearning system.

[1] http://moodle.org/
[2] http://www.blackboard.com/
[3] http://www.claroline.net/
[4] http://www.atutor.ca/

2.2 Improving eLearning web platforms through profiling

As their experience grows on using Web based eLearning platforms, educational institutions feel the need to improve their sites and related services. Reasons could be quite diverse. However, identifying, understanding and characterizing trends in what users do when they are navigating on a specific Web based eLearning site is with no doubt one of the most important reasons why institutions (and their webmasters and educational managers) are doing Web profiling nowadays (Lourenço & Belo 2009). Students using such platforms are a little bit different from the ones that access to non-educational Web sites. But, the way of acting is the same, as well are the tools and the services available. Thus, we can apply over Web based eLearning sites the same techniques that are currently used over any other site, independently from its business area, to catch the way of access, being, and explore a web site.

Web profiling (Spiliopoulou et al. 2000) is today one of the leading research and technical area of Web usage analysis (or clickstream analysis). The establishment of Web usage profiles allow us to evaluate the effectiveness of a site's contents, to improve their services performance and, consequently, to assess its popularity and effective use, which could help restructuring Web sites towards a more suitable platform for students and, of course, for their promoters.

There are several methods and techniques to improve Web sites organization and contents that we can use. Web usage analysis (Borges & Levene 1999) provides us the means to cover all the essential phases to discover a Web profile based on the page views request sequences records (clickstream data) we have available, namely (Ramadhan et al. 2005): pre-processing, where we read and interpret clickstream data in order to identify (if possible) an user (or a group of users) and its sessions, giving particular attention to information about applications, content server, and visited links and associated duration times; pattern discovery, in which we try to evaluate several functional and operational indicators such as sites more visited, sessions average lengths, IP addresses resolution, users' countries; most used agents, page views more frequent and respective access times, or navigations paths; and, finally, pattern analysis, where we deal with very specialized tasks that distinguish relevant information in clickstream data for user characterization. Classification and clustering, conventional statistics analysis and sophisticated visualization techniques tools are frequently used in the most Web usage analysis processes, helping a lot in the establishment of Web usage profiles. However, other techniques and models to identify profiles could be used.

3. Profiling with Markov chains

In Web Usage Mining (Srivastava et al. 2000), sequential patterns techniques (Esmaeili & Gabor 2010) try to find regular occurrences of a same set of elements (patterns) for each Web session under supervision, such as the presence of an item set followed by other item in a target group of page views. Using these techniques it is possible to predict future visit patterns that could be used to launch oriented advertisements, functional warnings, or simply to inform Web users about new things that they could be interested. Basically, we intend to find some relevant information about Web users behaviour, in order to provide more attractive site. A usual navigation process over an ordinary Web site begins by selecting an initial hyperlink and receiving the page view within a conventional Web browser. This first page view defines the initial state of a new Web session. Next, the navigation process continues selecting a new hyperlink that will define a new stage for the current Web session. This successive selection of hyperlinks is continuously recorded, click after click, in specific clickstream files. Latter, and with appropriated techniques, the analyses of these log files will reveal all the navigation paths (sequences of pages views) for a Web site that its users have done their sessions (Jansen et al. 2007).

As referred before, one of the most relevant approaches to map Web navigation paths is to use Markov chains (Sarukkai 2000). They were introduced as a mechanism to predict potential Web navigation links in a Web site (Zhu et al 2002) (Deshpande & Karypis 2004). With a Markov chain it is possible to indicate which is the next page will be requested by an user based on its current location and on its previous navigation sessions. Also, when generated based on Web transactions, the chains provide us the navigation paths followed by a user during a specific period of time, giving us the possibility to identify the most frequent page sequence that the user will probably follow in a future Web session (Borges & Levene 1999). Representing Web navigation paths with Markov chains we can get valuable information about the site in analysis and its future users navigation tendencies (Mobasher 2006) – for instance, through the analysis of a Markov chain, it is possible to predict what

pages view sequence a special kind of user will do in a newspaper, and prepare accordingly some advertisement spots that go towards the characterization of the navigation profile.

Markov chains are especially useful to built prediction models (Deshpande & Karypis 2002) (Surukkai 2000), allowing for the establishment of future user behaviour while users are interacting with the sites. This is done with the analysis of previously users behaviour with similar interests. We can use Markov models to find the more frequent trails (navigation paths) followed by users in their navigation processes, which means to find the most frequent sequences of pages that the users visit during their navigation sessions. Markov chains could be seen as a conventional graph in which nodes (stages) represent visited pages and edges (transitions) the probability related to some Web user passing from a page to another. Usually, edges' probabilities are calculated counting de number of users that pass from one page to another, taking into consideration the number of visits for each origin page. More formally, a Markov chain is characterized by a set of states {s1, s2, …sn} and a matrix of probabilities $[Pr_{ij}]_{nxm}$, where Pr_{ij} represents the probability to move from state i to state j.

The Markov chains are especially useful for predicting models based in continuous sequences of events – their application is quite direct to Web sessions. The order of a Markov model represents the number of prior states used to prevent the next state. So, a Markov model of order k predicts the provability of the next page based on the last k pages visited. Given a set of all trails R, the probability of reaching a state s_j from a state s_i, via a trail $r \subset R$, is given by $Pr(r) = \prod Pr_{k,k+1}$, where k ranges between i, and j-1, in other words the probability is given by the multiplication of all intermediate states (Mobasher 2005). As an example about how a markov chain can model a set of web transactions, consider the set of transactions presented in Table 1.

Table 1: An example of a transaction set

Id	Transactions
1	A → B → C → D → E
2	A → B → C
3	A → B → C → E
4	C → D → E
5	C → D → E → B
6	C → D → A → E
7	D → A → B → E

To build a Markov chain we start by adding an initial state (S) into the chain. This state will have a transaction for all the web pages visited for the target web site. Then we define a final state (F) that will be the end for every last page of either session. This means that every last page visited will have a transaction to this final state. The probabilities associated with the edges are obtained by counting the number of times that the transaction occurs in the trails. Thus, the probability to move from the initial state S to a state A that represents the page A is about 5/27 (0.18), where 5 is the number of times that the page A occurs, and 27 is the total number of requests. Using the same process, the probability to move from page A to page B is 4/5 (0.8), where 4 is the number of times that B occurs after A, and 5 is the number of times that A occurs. Finally, the probability to move from page E to the final state F is 5/6 (0.83), where 5 is the number of trails where E is the final state, and 6 is the number of times that E occurs. The Markov model generated from such transactions is depicted in figure 1.

However, a Markov chain it is not enough for a correct identification of a Web user profile. It is also necessary to calculate which are the most frequent Web paths (a sequence of Web links) followed on a specific site by its users. This is usually done after the Markov chain was built. To do that, we can use a common search algorithm for graphs, such as the *Breath-First Search* (Weiss 1993) or the *Depth-First Search* (Tarjan 1962) algorithms.

In order to calculate the frequent trails (table 2) we have to take into consideration two very important concepts: support and confidence. The support is given by the initial probability of the chain, which is frequently defined as the average of all the initial percentages of the Web pages, and the confidence by multiplying the probabilities of all links into the chain. The frequent trails will be calculated as follow: first we get all pages that have initial probability higher then the support value; then we use the

depth-first (Tarjan 1972) algorithm to search in graphs, trying to find the tails that have a probability bigger then the confidence.

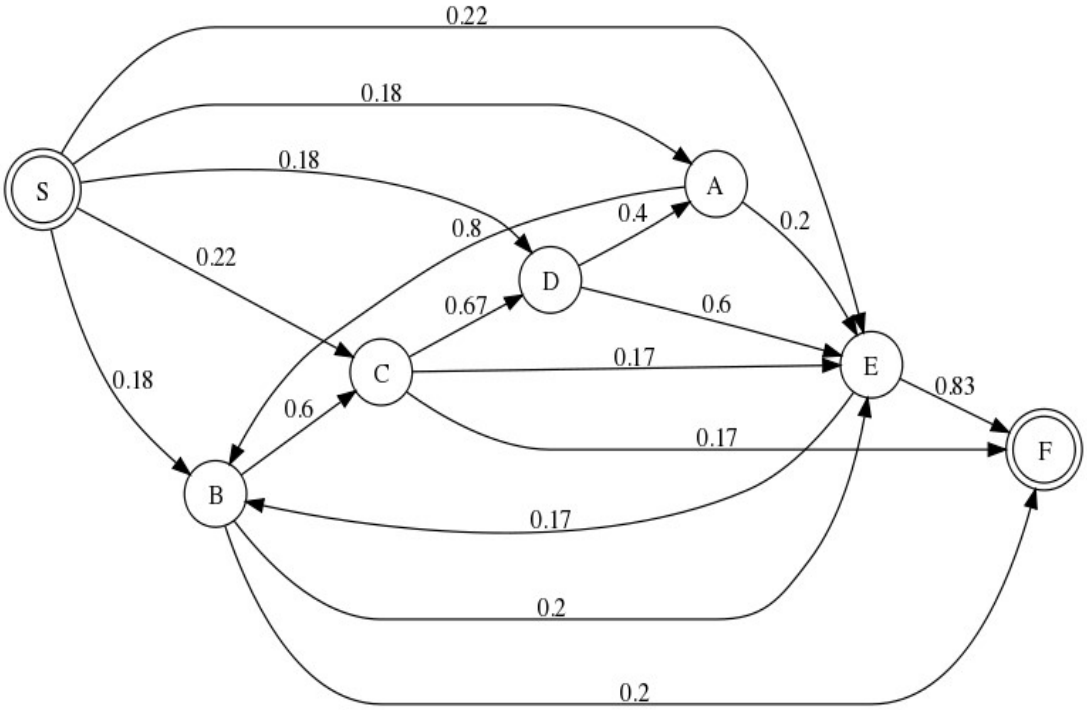

Figure 1: An example of a Markov Chain

Table 2: Frequent trails with support = 0.1 and confidence = 0.4

Trail	Probability
B -> C	0.6
B -> C -> D	0.4
D -> A	0.4
D -> E	0.6
A -> B	0.8
A -> B- > C	0.48
C -> D	0.67
C -> D -> E	0.4

Finally, analysing each identified path and using their own knowledge and expertise about the site itself, we can easily to establish the site's Web usage profiles, simply because paths characterize quite well the type of users that visit the site. These are only some general ideas about Markov chains and their utility in the establishment of Web usage profile - to have a more detailed idea about this we suggest the reading of (Sarukkai 2000), (Zhu et al 2002) or (Deshpande & Karypis 2004).

4. Establishing student web usage profiles

4.1 The eLearning site target

In order to demonstrate the utility of Markov chains in the establishment of user usage profiles in eLearning platforms, we selected one of our educational sites to be our target of study. This site (Figure 2) is supported by Moodle (Cole & Foster 2008) and maintains all the information about a MSc curricular unit that students need to be update about curricular topics, analyse and discuss practical issues, and receive or post their practical works. The eLearning site was structured to receive several courses. From its initial page, valid credential users can select and access to the material posted for each course on-line. The platform is able to support the main operations that ones need to use a site like this, namely: consult, insert, update or delete information items for all the resources available, communicate and exchange information with other students, follow the execution of practical works,

see evaluation processes, access and read technical reports and scientific papers, or participate on discussion areas. Additionally, we can also use some other management services that provide us generic information about current users and their activities.

The target public of our site are teachers and students of the courses the site maintain. As expected, the student community is the largest one. It is quite diverse and heterogeneous, and can be divided in different groups accordingly their own navigation credentials. Which means that it is a good target for our first studying process for discovering and establishing eLearning usage profiles.

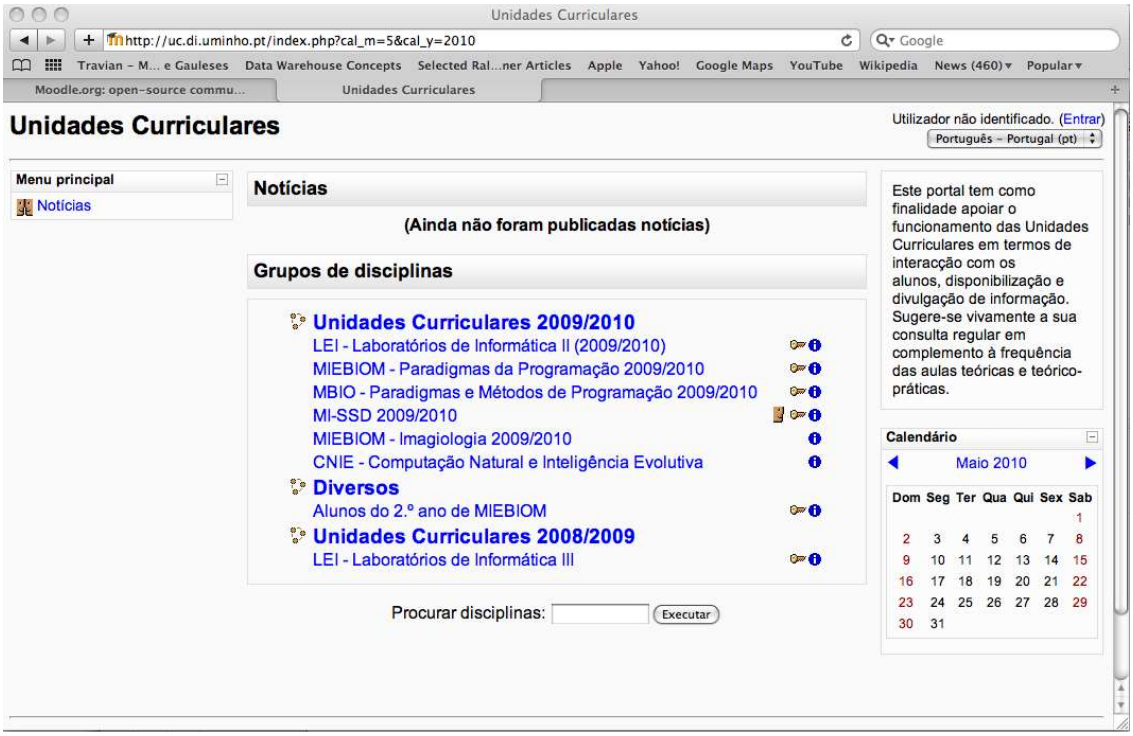

Figure 2: The target eLearning site

The site doesn't have a large volume of accesses, once it was designed and implemented for courses with small number of students. A preliminary site usage analysis gave us some curious information, but nothing very surprisingly to people that know a little bit about educational institutions and their working periods. Every time a course begins the site receive a lot of access requests and page viewing increases quite much when compared to other periods - students are looking for information about the functioning of the course, and preliminaries documents with course's program contents and evaluation criteria. A same high usage level happens again, every time one of the teachers post the results of a evaluation process or launch new tasks to do in the context of a specific working plan. These mean a very simple thing: every time students detect a new piece of information on the eLearning site they access it. Finally, and as expected, the majority of the accesses to our site were done from Portugal, once the students are Portuguese and every time they decide to visit the site use Portuguese access points.

Table 3: A simplified view of an event log file structure

Log File		
MI-SSD_1, 2009 Dezembro 10 19:19, 188.81.39.207, userName, course view, MI-SSD 2009/2010		
Field	Example	Description
Course	*Mi_SSD_1*	Course identification.
Date	*2009 Dezembro 10 19:19*	Date and time of the Web server when responding to a service request.
Host	*188.81.39.207*	User's IP address.
User	*UserName*	User name.
Page	*Course view*	Service requested.
Information	*MI-SSD 2009/2010*	Information about the service requested.

4.2 Sources and data preparation

All target data was collected in the information sources used by Moodle to identify and characterize students, and record their activities on site in a specific event log file. These log files are quite important to understand user behaviour. Their structure (Table 3) allows keeping detailed records about tasks users performed, including also date and time stamps and resources accessed.

Data preparation is one of the most important steps in profile identification and description. It requires that data will be transformed and clean according with the requirements of the data mining techniques to be used. As usual in these cases, to prepare data for profiling we need to perform several tasks that goes from users identification to data conciliation, passing for other tasks such as session delimitation, data cleaning and enrichment or IP addresses resolution, just to name a few.

We started the data preparation process by removing from the log file every line (record) that contained one or more null values in relevant attributes used on the users sessions reconstruction, because nulls do not let identify any kind of profile – as we know, a null value represents an unknown value and acts in any process as an absorbent element. Next we passed to the identification of the users that visit our site. This is quite important if we want to characterize their behaviour and consequently their usage profile. There are several methods we can use to make this identification (Cooley et al. 1999). However, in our case, this process was very simplified, once our target site demands that users identify themselves before access to the eLearning platform. So, when they do that, the system records their identification for any task they perform inside it. The next step was user session identification.

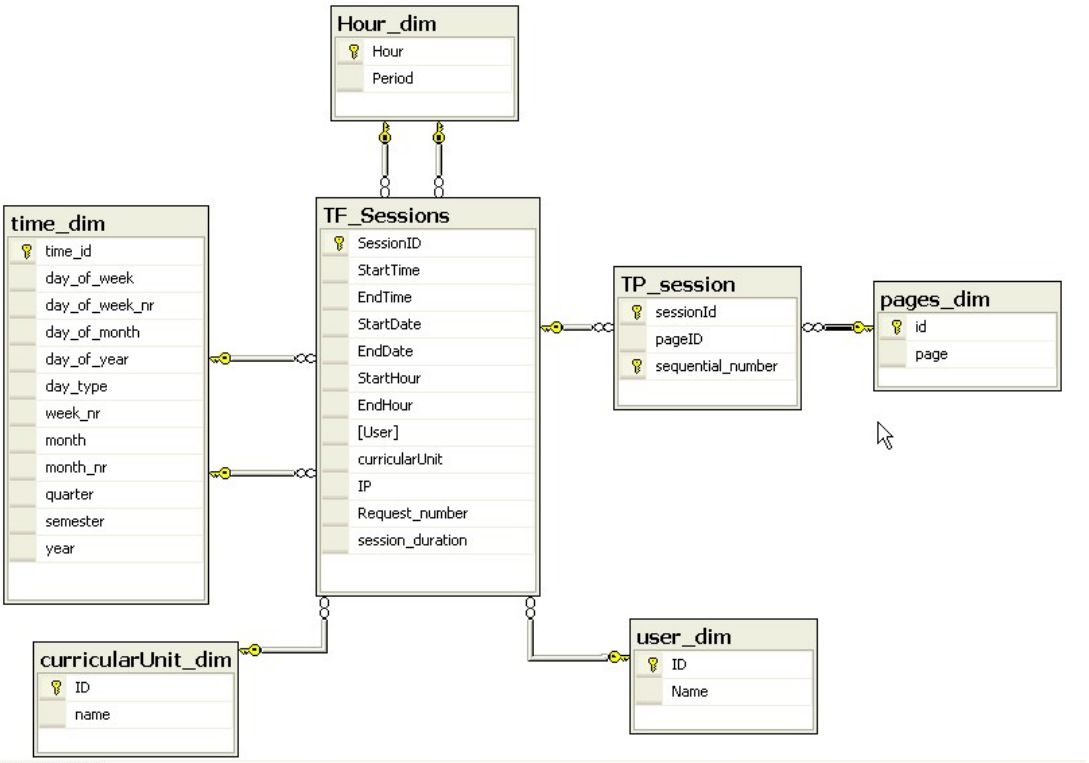

Figure 3: The user session star-schema

Basically, a session is a set of actions that a user performs in a site from the moment he enters it to the one he leaves it (W3C 1999). User session reconstruction intends to group under a unique identifier all the page views requests that a user did for a site. Keeping this kind of information (clickstream data) it is possible to study users' behaviour as well to know their navigation preferences and tendencies. We can divide session reconstruction techniques in two categories: proactive and reactive (Spiliopoulou et al. 2003). The first one considers that the session unique identifier is assigned during the navigation process, whilst the second defines it only when the session ends, appealing to the data recorded in the navigation logs. We used a reactive technique in our case study.

So, we began to define a maximum period of inactivity: 10 minutes. Next, we organized all page views requests by date and hour for each user registered in the site, in order to calculate user sessions. If two successive requests were posted in a time interval less than the maximum period of inactivity they belong to the same session. In the contrary case, we created a new session.

The last step in the data preparation process was loading all the prepared data in a data warehouse, which was especially designed and implemented according the requests for usage profile identification and characterization. The star-schemas organization followed the one presented in (Kimball & Merz 2000) for similar cases. We stored the data in two different star-schemas: one for receiving page views requests, and another to keep information about user sessions (figure 3). In this last star-schema we easily identify five dimension tables: 1) calendar (time_dim), which defines the different time periods (day_of_week, month, quarter, year, etc.) for user session analysis; 2) hour (hour_dim), corresponding to the different hours of the day and correspondent periods; 3) web pages (pages_dim) that contains the hyperlinks of every single page that was visited by users; 4) users (user_dim), which has a short identification of the user (only possible because all eLearning platform's services require access credentials); and curricular units (curricularUnit_dim), the dimension that has all the names of the curricular units supported by the platform. In order to define the N:N (many-to-many) relationship between the web pages dimension and the user session fact table (TF_Sessions) – one single Web session usually include several Web pages, and a Web page could be found in several distinct Web sessions - we created a bridge table (TP_Session). A quick look to this multidimensional schema allows us to see the enormous potential that a data structure likes this one has, and the enormous amount of queries that it is possible to satisfy about user sessions. For instance, if we want to know the average session time of all users that accessed the 'Computer Science' curricular unit site every 'Friday' during '2010', we can launch, in a simple manner, the following SQL query:

```
SELECT US.Name, AVG(TF.Session_Duration)
      FROM TF_Sessions AS TF INNER JOIN Time_Dim AS CA
            ON TF.StartTime = CA.Time_Id
            INNER JOIN User_Dim AS US
                  ON TF.[User] = US.Id
            INNER JOIN CurricularUnit_Dim AS CD
                  ON TF.CurricularUnit = CD.Id
      WHERE CA.Day_of_Week = 'Friday' AND CA.Year = '2010'
                  AND CD.Name = 'Computer Science'
      GROUP BY US.Name;
```

or to know the 10 most used curricular unit sites:

```
SELECT TOP 10 CD.Name, SUM(TF.Session_Duration)
      FROM TF_Sessions AS TF INNER JOIN CurricularUnit_Dim AS CD
            ON TF.CurricularUnit = CD.Id
      GROUP BY CD.Name
      ORDER BY SUM(TF.Session_Duration) DESC
```

or, finally, to get the most accessed Web page in all curricular unit sites:

```
SELECT TOP 1 PD.Page, COUNT(TF.Session_Id)
      FROM TF_Sessions AS TF INNER JOIN TP_Session AS BT
            ON TF.SessionId = BT.Session_Id
            INNER Pages_Dim AS PD
                  ON BT.Page_Id = BT.Id
      GROUP BY PD.Page
      ORDER BY COUNT(TF.Session_Id) DESC;
```

In this work we only used the data loaded in the user sessions star-schema, corresponding to 501 sessions performed by users between '2009/12/10' and '2010/05/26' and generated from a data set of 3585 requests. These requests involved 25 distinct users that accessed to 41 different pages. All this corresponds to the activity developed in an eLearning system for a single course integrated in a Decision Support Systems curricular unit. In spite of being a relative small data set, it was enough to

demonstrate the utility of Markov chains in the establishment of usage profiles – an excerpt of the contents of the user session fact table can be seen in figure 4.

4.3 The profiling process

After data preparation and integration, the next step on the profiling process is the generation of the Markov chains. This task begins with the reconstruction of the site's users sessions, which are necessary to support and represent the chain. To do that, we collected all the pages that were visited, and retrieved all the requests done over them, ordered by session and sequence number. With all the sessions reconstructed and all the visited pages stored, we calculated the initial probability for each visited page, dividing the number of requests for the page by the total number of requests. Having all the initial probabilities set, we must to calculate the probability that a page have to appear (be accessed) next to other.

	SessionID	StartTime	EndTime	StartDate	EndDate	StartHour	EndHour	Utilizador	Disciplina	IP	Pedidos	session_duration
1	1	2010-05-07 14:17:05.000	2010-05-07 14:17:05.000	2010-05-07	2010-05-07	14	14	1	2	NULL	1	0
2	2	2010-05-07 14:42:39.000	2010-05-07 14:50:24.000	2010-05-07	2010-05-07	14	14	1	2	NULL	8	465
3	3	2010-05-07 15:14:11.000	2010-05-07 15:14:56.000	2010-05-07	2010-05-07	15	15	1	2	NULL	3	45
4	4	2010-05-08 13:25:59.000	2010-05-08 13:29:28.000	2010-05-08	2010-05-08	13	13	1	2	NULL	6	209
5	5	2010-05-08 14:04:44.000	2010-05-08 14:14:21.000	2010-05-08	2010-05-08	14	14	1	2	NULL	14	577
6	6	2010-05-08 15:21:32.000	2010-05-08 15:21:46.000	2010-05-08	2010-05-08	15	15	1	2	NULL	2	14
7	7	2010-05-08 15:51:44.000	2010-05-08 15:51:50.000	2010-05-08	2010-05-08	15	15	1	2	NULL	2	6
8	8	2010-05-08 17:32:34.000	2010-05-08 17:32:37.000	2010-05-08	2010-05-08	17	17	1	2	NULL	2	3
9	9	2010-05-08 18:09:34.000	2010-05-08 18:20:30.000	2010-05-08	2010-05-08	18	18	1	2	NULL	16	656
10	10	2010-05-08 19:12:18.000	2010-05-08 19:14:18.000	2010-05-08	2010-05-08	19	19	1	2	NULL	9	120
11	11	2010-05-08 20:05:14.000	2010-05-08 20:10:36.000	2010-05-08	2010-05-08	20	20	1	2	NULL	12	322
12	12	2010-05-11 15:18:56.000	2010-05-11 15:18:56.000	2010-05-11	2010-05-11	15	15	1	2	NULL	1	0
13	13	2010-05-17 14:32:14.000	2010-05-17 14:32:14.000	2010-05-17	2010-05-17	14	14	1	2	NULL	1	0
14	14	2010-05-20 11:46:56.000	2010-05-20 11:51:28.000	2010-05-20	2010-05-20	11	11	1	2	NULL	7	272
15	15	2010-05-21 15:54:28.000	2010-05-21 16:44:56.000	2010-05-21	2010-05-21	15	16	1	2	NULL	17	3028
16	16	2010-05-23 15:01:34.000	2010-05-23 15:04:30.000	2010-05-23	2010-05-23	15	15	1	2	NULL	7	176

Figure 4: A record set fragment of the user session fact table

To do this, we need to get again all the pages visited, and page-by-page get all the sessions where each of them appears. Ending that, it was necessary to find what page appears next to the previous referred page in all sessions. If that happens for a first time we initialize a counter with 1; If not, we increment the counter by 1. At the end we have an array with all the pages next to the reference page and the number of times they appeared in that position. The probability of the next page is calculated taking the number of times that a page appeared next to the reference page and dividing it by the number of times that the reference page appears. In Figure 5 it is presented an excerpt of the Markov chain generated for the site, using the tool GraphViz[5] - due to the large number of nodes and edges of the Markov chain generated, we only present part of it in this paper.

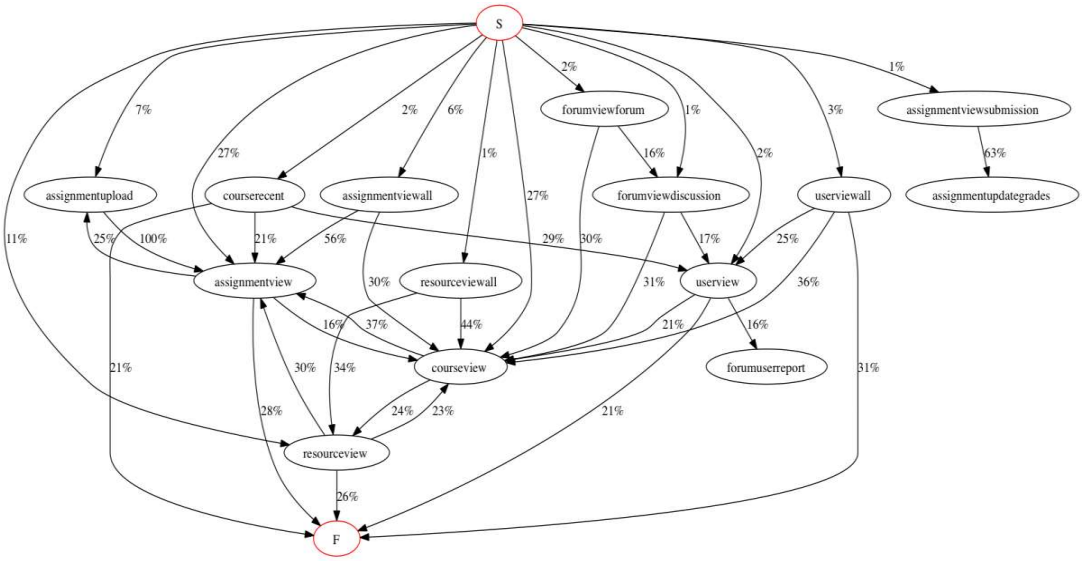

Figure 5: An excerpt of the Markov chain generated for the site

[5] http://www.graphviz.org/

After the generation of the Markov chain and getting the user accesses data, we calculated the most frequent paths followed by users. Then, we defined the minimum support, using the average value of all the initial probabilities defined previously, and the confidence, selecting a value of 15% due to the fact that users had a large number of pages for selection. If we had chosen a greater value for the confidence, an important number of chains simply was despised and not included in the final results. To get the most frequent paths we need to: retrieve all the pages views that have a probability greater than the support that was defined; calculate all the paths whose confidence is greater that the defined value using the *depth-first algorithm for graphs* (Tarjan 1962); and, finally, show a list (Table 3) of the most frequent paths with a support of 1% and a confidence of 15%. At this time, these are two important concepts to consider - support, represents the initial provability of a Web path, many times defined as the average of the initial percentages of all the Web pages; and the confidence, corresponds to the probability of an user going on a specific path, being calculated by multiplying the probabilities of each chain connections.

Through the analysis of the most frequents paths that were detected it was possible to identify some relevant site usage profiles. At the beginning, we had made the identification of two distinct groups of users that access regularly to the Web based eLearning site: teachers and students. It was a simple observation of a fact. However, if we look now to the rules that were generated (Table 3), we can see that paths 1, 7, 12 and 19 are associated with the teacher profile, since they include actions related to the creation of new tasks on the site, which are obviously from its competence. Looking for the other rules, we can say that they are typical from a student usage profile. Rules 3, 4, 5, 7, 10, 14, 15, 16, 20 and 22 reveal a student profile a little bit more specific. This first student profile corresponds to a group of users that only visit the site to see information about their curricular units and the tasks they have to do. There is a second student profile that was identified and rules number 6, 11,13 and 18 support it corresponding to users that do the same as the previous profile but also visit the pages from other users defined in the site. Finally, rules 8, 9, 18, 20 and 21 reveal a third students group that are more active on the site, participating in the discussion lists besides other regular activities of studying and working tasks.

Table 3: A list of the most frequent paths

Id	Path	Probability
1	assignmentupload -> assignmentview	100
2	assignmentviewsubmission -> assignmentupdategrades	63
3	assignmentviewall -> assignmentview	56
4	resourceviewall -> courseview	44
Id	Path	Probability
5	Courseview -> assignmentview	37
6	userviewall -> courseview	36
7	resourceviewall -> resourceview	34
8	forumviewdiscussion -> courseview	31
9	forumviewforum -> courseview	30
10	resourceview -> assignmentview	30
11	courserecent -> userview	29
12	assignmentview -> assignmentupload	25
13	userviewall -> userview	25
14	Courseview -> resourceview	24
15	resourceview -> courseview	23
15	resourceview -> courseview	23
16	courserecent -> assignmentview	21
17	userview -> courseview	21
18	forumviewdiscussion -> userview	17
19	assignmentupload -> assignmentview -> courseview	16
20	resourceviewall -> courseview -> assignmentview	16
21	forumviewforum -> forumviewdiscussion	16
22	assignmentview -> courseview	16
(…)	(…)	(…)

5. Conclusions and future work

In this paper it was presented a first attempt to establish valid and useful students usage profiles for a Web-based eLearning system of an educational institution. As referred previously, the main goal was to understand how students used the resources that teachers made available to them in an eLearning site. Nowadays, the success of any eLearning platform is directly dependent from the knowledge that its administrator has about their users. Usage profiles can be use by administrators to personalize contents and services towards their users' needs and expectations. We can use several data mining techniques to identify usage profiles like clusters, association rules, sequential patterns, classification, and so on. The goal of this paper was to study the application of Markov chain to identify student usage profiles of a specific Web based eLearning platform. In general terms, we started by a brief explanation about how Markov chains can be used to discovery web usage profiles, and then we present a case of study where we used Markov chains as a way to identify and establish usage profiles. The intention was to get valid profile information in order to optimize the site's structure and reduce operational costs involved with the maintenance of resources not used. The results obtained were quite typical for a system like the selected one. We identified two large user groups (teachers and students) whose behaviour was divided, respectively, doing creation and maintenance tasks over the site's contents, and consulting working tasks or participating in discussion lists. Doing this, it was possible to demonstrate the establishment of real usage profiles of the target site through the use of Markov chains. If we had a more significant amount of log records, results will be more remarkable revealing other potential profiles beyond the ones that we obviously expect. This will be explored soon in one of our research lines, especially oriented to the refinement of the generate Markov chains, filtering irrelevant access to some pages in order to discover the most accessed sites and correspondent visit times. This will improve our perception about eLearning sites usage, contributing to improve sites' organization, contents, interface and to reduce computational costs related to pages that never will be consulted.

References

Borges, J., Levene, M. (1999), Data Mining of User Navigation Patterns. In of the Web Usage Analysis and User Proling Workshop, pages 31–36, San Diego, USA, 1999.

Cole, J., Foster, H. (2008), Using Moodle, 2nd Edition, O'Reilly Media, Inc.

Cooley, B., Mobasher, B., Srivastava, J., (1999), "Data preparation for mining World Wide Web browsing patterns". Journal of Knowledge and Information Systems, 1(1), 1999.

Carliner, S., Shank, P. (2008), The ELearning Handbook: A Comprehensive Guide to Online Learning, Pfeiffer, April 25.

Deshpande M., Karypis G. (2004), Selective Markov Models for Predicting Web Page Accesses, ACM Trans. on Internet Technology.

Esmaeili, M., Gabor, F. (2010), Finding Sequential Patterns from Large Sequence Data. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January.

Hosam F. El-Sofany, Ahmad M. Hasnah, Jihad M. Jaam and Fayed F. M. Ghaleb., "A Web- Based E-Learning System Experiment". Proc. of the Intel. Conf. on E-Business and E-learning, PSUT, Amman-Jordan, 112-119, 2005.

Kimball, R., Merz, R. (2000), The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. John Wiley & Sons, Inc.

Jansen, B. J., Spink, A., Blakely, C., and Koshman, S. (2007), Defining a session on Web search engines: Research Articles. J. Am. Soc. Inf. Sci. Technol. 58, 6 (Apr. 2007), 862-871.

Lourenço, A., Belo, O. (2009), Web Crawler Profiling and Containment Trough Navigation Pattern Mining, in Proceedings of the IADIS WWW/Internet 2009, 19-22 November, Rome, Italy.

Mobasher, Web Usage Mining and Personalization. In Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press. 2005

Mobasher, B. (2006), Web Usage Mining (Invited Chapter). Web Data Mining: Exploring Hyperlinks, Contents and Usage Data (by Bing Liu). Springer Berlin-Heidelberg.

Ramadhan, H., Hatem, M., Al-Khanjri, Z., Kutti, S. (2005), A Classification of Techniques for Web Usage Analysis. Journal of Computer Science, July.

Rokou F., Rokou, E., Rokos, Y., (2004) "Modeling web-based educational systems: process design teaching model," Educational Technology and Society, Vol. 7, pp. 42-50.

Sarukkai R. (2000), Link Prediction and Path Analysis Using Markov Chains. In Proceedings of the 9th Intl.World Wide Web Conference.

Schewe, K., Thalheim, B., Binemann-Zdanowicz, A., Kaschek, R., Kuss, T., and Tschiedel, B. (2005), A Conceptual View of Web-Based ELearning Systems, Education and Information Technologies 10:1/2, 81–108, Springer.

Spiliopoulou, M., Pohle, C. and Faulstich, L.C. (2000). Improving the Effectiveness of a Web Site with Web Usage Mining. Lecture Notes in Computer Science, 142-162.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD Explorations Newsletter, 1(2), 12-23.

Tucker, S., Pigou, A., Zaugg, T. (2002), "eLearning Making It Happen Now," in Proceedings of 30th Annual ACM SIGUCCS Conference on User Services, 2002, pp. 292-293.

Tarjan R. (1962), "Depth-first search and linear graph algorithms". Siam Journal on Computing.

W3C (1999), World Wide Web Commitee Web Usage Characterization Activity: "W3C Working Draft: Web Characterization Terminology & Defenitions Sheet". 1999. http://www.w3c.org/1999/05/WCA-terms/.

Weiss M. A. (1993), " Data Structures and Algorithm Analysis in C". The Benjamin/Cummings Publishing Company, Redwood City, Califórnia.

Zhu, J., Hong, J., Hughes, J. (2002), Using Markov Models for Web Site Link Prediction. Proceedings of the 13th ACM conference on Hypertext and Hypermedia (Hypertext'02), pp. 169-170, ACM Press, College Park, MD, USA, June 11-15.

