# User navigational behavior in e-learning virtual environments

Joan Marc Carbó, Enric Mor, Julià Minguillón
Computer Science and Multimedia Studies
Universitat Oberta de Catalunya
08035 Barcelona, Spain
{jmcarbo, emor, jminguillona}@uoc.edu

## Abstract

*In this paper we describe the navigational behavior of the students of a e-learning virtual environment, in order to determine whether such navigational patterns are related to the academic performance achieved by the students or not, and which behaviors can be identified as more successful. As an example, a subset of students taking a degree in Computer Science in a completely virtual online university are selected as the matter of study. Three levels of analysis are described: a session level, where students perform a few actions in a single session logged to the virtual campus; a course level, where all single sessions are joined to form a course navigational pattern; and a lifelong learning level, where students enroll in several subjects each academic semester. A simple experiment is outlined for the course level to demonstrate the possibilities of such analysis in a virtual e-learning environment. This experiment shows that the information collected in this level is useful for understanding user behavior and the relationship with his or her academic achievements, and that some intuitive ideas about the relevance of specific user actions or particularities can be also better explained.*

## 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 information in any web-based organization system is described in several papers, see [13, 7] for example. There are several fields where web mining can be used for understanding user navigational behavior. This expertise about users behavior can be reintegrated in the web-based system (offering user personalized services, for example) in order to improve user experience and satisfaction, and hopefully, to strength the customer relationship model.

On the other hand, e-learning is one of the most promising and growing issues in the information society nowadays. 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 appear of Learning Management Systems (LMS) 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 [11] 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 users behavior, navigational patterns and other issues related to the learning process can be formulated and validated by means of web mining tools.

This paper is structured as follows: Section 2 describes the internal structure of e-learning virtual environments and the navigational levels. The experiments which support the user navigational behavior hypothesis are performed and discussed in Section 3. Finally, in Section 4, the current and future research lines in this subject are outlined.

Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)
0-7695-2415-X/05 $20.00 © 2005 **IEEE**

## 2. E-learning virtual environments

The intensive use of Internet possibilities for both content searching and delivery but also for interface design and implementation has completely changed the visions in the 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 learning, the government and other sectors. 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 [6] may be used to perform an automatic 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 but also 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 [10] environment: monitoring and interpreting the activities of its users [16], inferring user requirements and preferences, and acting upon the available knowledge on its users.

### 2.1. The UOC virtual campus

The Universitat Oberta de Catalunya [12] (UOC[1], in English known as Open University of Catalonia) is a completely virtual campus which offers 19 official degrees, several graduate programs and post-graduate studies, and a doctoral degree, with more than 35000 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 e-learning related 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 [1] is becoming

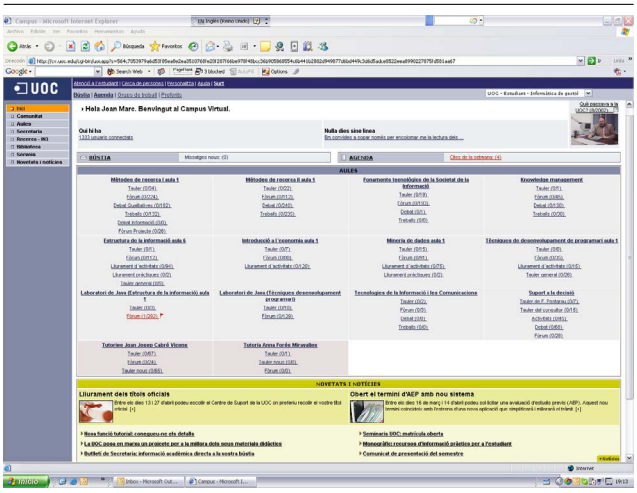

**Figure 1. Virtual campus start page once user has logged in.**

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, that is, to successfully achieve the learning goals and fulfil the course requirements in order to pass

1 http://www.uoc.edu

the course exams. This fact can be also used to determine the variables used in the clustering process. 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 satisfaction by means of personalization [8]. The inclusion of e-learning standards will allow a better tracking of students behavior (using the SCORM 2004 standard [1] 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 management enquiries service; and interaction with professors and other students through pre-defined communication channels (e.g., forums, 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 modelling purposes. Most of this information is collected by the web servers in form of server log files, according to the Apache Common Log Format.

### 2.3. Server log files

For each action a user performs in the virtual campus, one or more lines representing such action 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 them have been joined in a single file which is generated each day. This file is firstly preprocessed 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 preprocessing 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 lines, approximately 12 GB, which is a very large figure. Therefore, a second preprocessing, more oriented towards narrowing the experiment, is required.

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.

### 2.4. 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 (that is, 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 to the proposed exercises. At this level, the short-term navigation

behavior is studied, that is, what each individual user makes 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, for example [5].

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. 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 is 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 an extensive usage of the virtual classroom forums during the weekend or not. All the information collected at this level could be used to feed an intelligent tutoring or adaptive hypermedia system (see [15, 3], for example), with personalization purposes.

The third level, namely the lifelong learning level, it 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 (see [14], for example). At this level it may be interesting to discover unappropriate 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.

## 3. 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. A subset of 111 students taking a degree in Computer Science, which are enrolled in the subject "Compilers I", has been selected as the matter of study. These students may also be enrolled in other subjects, but we will focus in the navigational patterns related to the subject matter of study. Other experiments including this and other data sets can be found in [9].

| Mark | A | B | C+ | C- | D | N |
|---|---|---|---|---|---|---|
| Total | 22 | 65 | 12 | 1 | 0 | 11 |

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

Basically, students connect to the virtual campus and access to the virtual classroom to follow the learning activities designed for each subject, according to a previously established scheduling. In the case of the subject "Compilers I", students are asked to solve an optional exercise which is published during the second week (once the course has been started) and that it must be solved and returned back to the teacher within 12 days (including two weekends). 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 to mention 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 do a final exam at the end of the semester, which has usually a higher degree of difficulty. We are interested in studying which students decide not to do the first exercise (or 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 could be collected online by an intelligent tutoring system and help each student to 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 to mention that it is a simple exercise for introducing them into the compilers subject, with a medium degree of difficulty, so it is not expected to have many students with poor marks, that is, 'C-' or 'D', but on the contrary, more students not doing it, that is, 'N'). In fact, eleven students decide not to do the proposed exercise, and only one does it poorly and fails.

### 3.1. Data preprocessing 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 "Compilers I". The final log file has 220000 lines approximately, which is a reasonable data set

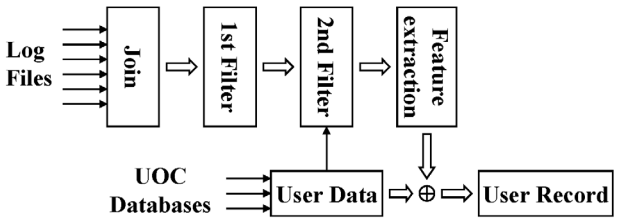


**Figure 2. Preprocessing steps for obtaining user records.**

for studying user navigational behavior in a focused environment. Figure 2 shows all the preprocessing steps, starting from the original log files and databases, until the final data set is generated.

The main advantage of being into 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 amount of subjects he or she is enrolled to, the number of semesters he or she has been studying in the UOC, and so. Other information such as the number of book loans requested or user satisfaction surveys are also available. Although this study may arise several questions about privacy issues (see [4], for example), 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. Therefore, we will use the following variables as the input for the classification and clustering algorithms:

- Gender (94 Male, 17 Female). 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 (ranging from 22 to 50 years old). This variable might provide important information about the sociodemographical background of students, as older people may have family burden, for example. If so, this information could be used to personalize the learning process according to such knowledge.
- Whether the student is new to the UOC virtual campus or not (13 are, 98 are not). New students may experience difficulties using the virtual campus, so it is interesting to validate such hypothesis.
- Whether the student takes the course for first time or not (82 do, 29 do 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.

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, the mean delay between two consecutive sessions, the mean length of each session, and the mean number of hits (user-driven actions) in each session. 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.
- A set of information related to the course level: the number of messages written in the appropriate virtual classroom forum, and the delay between the moment that the proposed exercise is published and the moment student accesses to its content.

Thus, 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. 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. It is worth to mention 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.

### 3.2. 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 [17] and supervised classification by means of classification and regression trees [2] 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. The TwoStep algorithm tries 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. This first study tries to identify which variables are relevant for classification purposes. A posterior supervised classification analysis could be used to design a recommendation system or an adaptive system for tutoring purposes. The experiment is designed to force the TwoStep algorithm to produce clusters which separate the students

who render the proposed exercise from those who do not. The optimal clustering is obtained with two clusters, as expected. The first cluster is formed by the 100 students who render the proposed exercise, while the second cluster is formed by the other 11 students who do not render it. Surprisingly, the student who renders a poor exercise and fails has a navigational behavior more similar to those who successfully solve the exercise than to those who decide not to do it.

This study is also useful to determine which variables are relevant and which are not for clustering purposes. As expected, both gender and age are not relevant for clustering purposes, and a statistical significance analysis reveals that no difference is found between the members of the two found clusters. The other variables describing student background are also not relevant, which may indicate two facts: first, students that are new to the virtual campus do not obtain worse results than the others, what can be mainly explained with two reasons: a) they are computer science students, which obviously reduces the digital gap among them and b) all students have attended a short course about virtual campus services and digital literacy with the same aim. Therefore, they are supposed to begin the course in the same conditions with respect to the virtual campus usage. Second, students that are enrolled into the subject for second time are in fact doing the activities as if it was the first time, mainly because they probably dropped out at the beginning of the previous course. These hypotheses should be posteriorly confirmed by means of a survey, in order to detect any problem or issue that could be useful for improving the subject learning process.

On the other hand, there are several variables which are indeed relevant for clustering. The most relevant ones are the mean delay between consecutive sessions (significative for $p < 0.001$) and the number of messages posted in the virtual classroom during the studied period of time ($p < 0.001$). In the first case, this variable is related to the total number of sessions, although it needs further explanation using a more detailed analysis. The second case is much more interesting: although students are not obliged to participate in the forums of the virtual classroom, the act of posting a message may mean a higher degree of interest and commitment to the subject. Message content is usually one of the following three kinds: for introducing themselves, asking for help about the proposed exercises, or answering the questions of other students. In fact, students may ask the teacher through the private mail and thus not use the virtual classroom space, but teachers are told to redirect students to such space in order to increase the interaction between students and share the knowledge created during the learning process. Other relevant variables are the mean length of each session ($p < 0.01$), which is longer for students who render the exercise than for the others, and the gap between the publication of the exercise and the moment that the student downloads it ($p < 0.1$), which is smaller for the successful students.

## 4. Conclusions

In this paper 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 a subject in computer science, namely "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. Although the results shown in this paper 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.

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.

## Acknowledgements

This work is partially supported by Spanish government grant MULTIMARK TIC2003-08064-C04-04.

## References

[1] A. D. L. (ADL). Sharable Content Object Reference Model (SCORM) 2004 2nd edition overview, 2004.

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

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

[4] C. Clifton and V. Estivill-Castro, editors. *Proc. of the ICDM 2002, Workshop on Privacy, Security and Data Mining*, volume 14, Maebashi City, Japan, 2002. ACS.

[5] M. Ivory and M. Hearst. The state of the art in automated usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):173–197, Dec. 2001.

[6] J. Kay and A. Lum. Creating user models from web logs. In *Proceedings of the Intelligent User Interfaces Workshop: Behavior-Based User Interface Customization*, Jan. 2004.

[7] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM*, 2, 2000.

[8] E. Mor and J. Minguillón. E-learning personalization based on itineraries and long-term navigational behavior. In *Proceedings of the Thirteenth World Wide Web Conference*, volume 2, pages 264–265, New York City, NY, May 2004.

[9] E. Mor, J. Minguillón, and J. M. Carbó. *Data Mining in E-learning*, chapter Analysis of User Navigational Behavior for E-Learning Personalization. WIT Press, 2005. (To appear).

[10] A. Paramythis and S. Loidl-Reisinger. Adaptive learning environments and e-learning standards. *Electronic Journal of e-Learning*, 2(2), Dec. 2004.

[11] D. Riecken. Personalized views of personalization. *Communications of the ACM*, 43(8):27–28, Aug. 2000.

[12] A. Sangrà. 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), Jan. 2002.

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

[14] C. Tattersall, B. van den Berg, R. van Es, J. Janssen, J. Manderveld, and R. Koper. Swarm-based adaptation: Wayfinding support for lifelong learners. In *Proceedings of the Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 3137 of *Lecture Notes in Computer Science*, pages 336–339, Eindhoven, The Netherlands, Aug. 2004.

[15] J. M. Tchétagni and R. Nkambou. Hierarchical representation and evaluation of the student in an intelligent tutoring system. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, volume 2363 of *Lecture Notes in Computer Science*, pages 708–717, Biarritz, France and San Sebastián, Spain, June 2002. Springer.

[16] R. Thomas, G. Kennedy, S. Draper, R. Mancy, M. Crease, H. Evans, and P. Gray. Generic usage monitoring of programming students. In *Proceedings of the ASCILITE 2003 Conference*, pages 715–719, Adelaide, Australia, Dec. 2003.

[17] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103–114, Montreal, Canada, 1996.