This document is published in:

International Journal of Technology Enhanced Learning (2012), 4 (1/2), 99-120.
DOI: http://dx.doi.org/10.1504/IJTEL.2012.048313

Peeking into the black box: visualising learning activities

Raquel Crespo García, Abelardo Pardo, Carlos Delgado Kloos

Department of Telematic Engineering, Carlos III University of Madrid, Spain,
E-mails: {rcrespo, abel, cdk}@it.uc3m.es

Katja Niemann, Maren Scheffel, Martin Wolpers

Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven,
53754 Sankt Augustin, Germany
E-mails: {katja.niemann, maren.scheffel, martin.wolpers}@fit.fraunhofer.de

Abstract: Learning analytics has emerged as the discipline that fosters the learning process based on monitored data. As learning is a complex process that is not limited to a single environment, it benefits from a holistic approach where events in different contexts and settings are observed and combined. This work proposes an approach to increase this coverage. Detailed information is obtained by combining logs from a LMS and events recorded with a virtual machine given to the students. A set of visualisations is then derived from the collected events showing previously hidden aspects of an experience that can be shown to the teaching staff for their consideration. The visualisations presented focus on different learning outcomes, such as self learning, use of industrial tools, time management, information retrieval, collaboration, etc. Depending on the information to convey, different types of visualisations are considered, ranging from graphs to starbusts, from scatter plots to heatmaps.

Keywords: learning analytics; learning activity visualisation; educational datasets; experimental datasets; learning activity; visualisation; data visualisation; student monitoring; attention metadata; observation; virtual machine; technology-enhanced learning; datatel.

Reference to this paper should be made as follows: Crespo García, R., Pardo, A., Delgado Kloos, C., Niemann, K., Scheffel, M., Wolpers, M. (2012) 'Peeking into the black box: visualising learning activities', *Int. J. of Technology Enhanced Learning* , Vol. ?, Nos. ?/?, pp.???–???.

Biographical notes: Raquel Crespo García received her M.Sc. degree in Telecommunications Engineering from the Technical University of Madrid, in 1999, and the Ph.D. degree on Communications Technologies

from the University Carlos III of Madrid in 2007. She is Associate Professor of Telematics Engineering at the University Carlos III of Madrid. She has been involved in a number of European, national and regional research projects. Her current research interests are in technology enhanced learning, adaptation, intelligent tutoring systems, computer supported collaborative learning, learning analytics and educational data mining.

Abelardo Pardo received his M.Sc. degree in Computer Science from the Polytechnic University of Catalonia, Spain, in 1991, and the Ph.D. degree in Computer Science from the University of Colorado at Boulder in 1997. He is an Associate Professor of Telematics Engineering at the University Carlos III of Madrid. His current research interests are in technology enhanced learning, intelligent tutors, adaptation, computer supported collaborative learning, learning analytics and technology enhanced tutoring strategies.

Carlos Delgado Kloos received the Ph.D. degree in Computer Science from the Technical University of Munich and in Telecommunications Engineering from the Technical University of Madrid. He is Full Professor of Telematics Engineering at the University Carlos III of Madrid, where he is the director of the Nokia Chair (www.it.uc3m.es/nokia/) and of the GAST research group (www.gast.it.uc3m.es). He is also Vice-Rector of Infrastructures and Environment. His main interests include Internet-based applications and in particular e-learning. He has been involved in more than 20 projects with European (Esprit, IST, @LIS, eContentPlus), national (Spanish Ministry, Region of Madrid) and bilateral (Spanish-German, Spanish-French) funding. He has published more than 200 articles in national and international conferences and journals. He has further written a book and co-edited five. He is the Spanish representative at IFIP TC3 on Education and Senior Member of IEEE.

Katja Niemann studied computational linguistics and computer science at the University of Heidelberg and received her master's degree in 2007. Since 2008, she has been working as a research associate at the Fraunhofer Institute for Applied Information Technology FIT and participated in several public projects in the domain of Technology Enhanced Learning. Her research interests are in Information Retrieval, Recommender Systems, Attention Metadata, and User Support.

Maren Scheffel studied at the University of Edinburgh, UK, and the University of Bonn, Germany, where she received an M.A. in Computational Linguistics. She currently works as a research associate at the Fraunhofer Institute for Applied Information Technology FIT. As a member of the "Context and Attention in Personalised Learning Environments" department she works on the FP7 IP ROLE (Responsive Open Learning Environments) where she is involved in the project's management. Her research focuses on the application of linguistic methodologies to the analysis of usage data.

Prof. Dr. Martin Wolpers holds a PhD in electrical engineering and information technology from the Leibnitz University Hannover. He is head of the research department "Context and Attention in Personalised Learning Environments" at the Fraunhofer Institute for Applied Information Technology FIT in Sankt Augustin, Germany. He also holds the title "visiting professor" at the Katholieke Universiteit Leuven, Belgium. He is strongly involved in a number of highly

successful European Projects. He is vice-president of the European Association of Technology Enhanced Learning (EATEL) and president of the MACE Association. His research interests focuses on the improvement of Technology Enhanced Learning by relying on new and emerging technologies. This includes work on knowledge representation and reasoning, knowledge management and provision as well as contextualised attention metadata and information elicitation (among others).

1 Introduction and Background

Most higher education institutions use learning management systems (henceforth simply LMS) for students to interact with learning resources, forums, wikis, blogs and other services (Ertugrul, 2000). These platforms usually record all the events that take place in so called "log files" that are accessible to administrators. In recent times, "Academic Analytics" has been described as the application of business intelligence technology to a learning context (Goldstein and Katz, 2005). Even more recently, the area of "Learning Analytics" has emerged where the events observed are used to improve a learning experience (Long and Siemens, 2011). There are numerous initiatives that analyse these events, detect anomalies in a learning scenario, and prescribe remediations. For example, the Signals project detects students at risk of dropping out of a course (Arnold, 2010; Tanes et al., 2011) from their events in the LMS. There are various interesting features about a learning experience that can be derived from the analysis of forum posts, homework submission patterns (Lin et al., 2009; Mazza and Dimitrova, 2004). But the LMS captures only a subset of all the events that occur in a learning experience. This subset is even smaller in blended learning or face to face learning scenarios. Considering data only from a LMS offers what could be called a "LMS-centric" approach. If a learning experience approaches a Web 2.0 paradigm, there are activities that are hosted in external applications (Chatti et al., 2007). Also, in experimental science courses it is highly common for activities to include procedures that are exercised outside of the LMS (Auinger et al., 2009). Finally, with the abundance of information available on the Internet, event if a course has a comprehensive set of resources, students are likely to access material outside the institutional LMS (Waycott et al., 2010).

Another research area that deals with user observations within learning environments is intelligent tutoring systems (Woolf, 2008). These applications are designed to observe and react to user actions. Typically, they guide the student through a learning process adapting the content, the activities and their navigation structure based on a user model and a knowledge domain model. However, these systems are typically highly specialised and therefore have the same drawback as the LMSs: there are valuable observations that occur outside the scope of the system. This observations point to the need of more exhaustive mechanisms to collect data about the events occurring in a learning environment.

The raw monitored data requires appropriate processing and analysis for the user to exploit them. Visualisations are expected to allow users to understand

and discover patterns in data more easily (Govaerts et al., 2010) and are thus a common approach. In the learning domain there exist several tools that allows visualising the students' monitored activity, most of them focused on online environments. The Student Analysis Monitor (Govaerts et al., 2010) provides a set of visualisations of learner activities regarding the use of resources in a PLE to increase awareness and to support self-reflection. The Participation Tool (Janssen et al., 2007) visualises the online communication of students doing group work. CourseVis (Mazza and Dimitrova, 2004) support instructors in web-based distance education visualising interaction in discussions, quiz performance and page accesses based on LMS logs. Moodog (Zhang et al., 2007) is a Moodle plug-in that visualises data from the activity logs to allow students to compare their progress to others and teachers get insight into the student interactions with the online course, and use of resources. Regarding offline activity, CAMera (Schmitz et al., 2009) supports personal monitoring and reporting, providing simple metrics, statistics and visualisations of the activities of the learner captured using Contextualised Attention Metadata (CAM) (Wolpers et al., 2007; Schmitz et al., 2007). Although this work is inspired on CAMera and also based on CAM for interoperability, it provides a wider set of visualisations and higher level indicators. Besides, our work focuses on more experimental-based learning activities (e.g. CAMera evaluation reports its use by researchers and PhD students, monitoring activities such as paper writing and mail-based collaboration).

A comprehensive set of observations used to analyse and derive concrete actions to intervene in a learning environment may benefit two target users. The first target group are the students. The observed events offer the opportunity of reflecting on how effective is their work. According to the literature, visualisations of learning activities are expected to motivate students, provide feedback on their work, facilitate collaboration, improve awareness and support self-reflection. The second group are the instructors. By accessing the collected data, instructors may complement their knowledge of how students are performing, how is learning taking place and detect strong and weak points of the entire experience. The approach presented in this paper addresses this second target group.

In the described context, monitoring students while they learn is a quite challenging task. Interaction with learning activities can be intermixed with other tasks not related to the learning scenario. There are a variety of tools emerging that allow a close monitoring of all the operations that occur in a personal computer (for example `wakoopa.com`, or `rescuetime.com`). But these products record all the events occurring in a computer either related to a learning activity or not. The desirable approach is to collect only those events related to the participation on a learning experience.

In this paper a set of visualisations of student events occurring outside a Learning Management System is presented. Data is collected by providing students with a virtual machine fully configured with all the required tools for a specific learning experience. The virtual machine is ready-to-use for the students to develop all required personal and team work inside it, covering both offline (editing, developing, etc.) as well as online (e.g. browsing) activities. The proposed visualisations have been conceived to offer instructors information about those aspects that escape their supervision. The application and visualisations have been used in a second year engineering course.

We aim to increase the observation capacity of the teacher, broadening the scope of observable student activities. We aim to boost the visibility of learning events, widening the monitoring for reaching scenarios traditionally opaque for the teacher, and thus perceived as a "black box". This information is complementary to more common learning analytics settings, like previously mentioned LMS logs. Combining the information from multiple sources will provide a more complete and accurate vision of the learning process. In consequence, we aim to analyse scenarios typical in a course, such as:

- Students' participation and interactions in the course forum
- Students' browsing activity outside the LMS: online resources they check, sites they visit, etc.
- Students' work outside the LMS: what kinds of activities do student perform? Apart from browsing the LMS contents, students are expected to work on their own, typically offline, editing documents, programming, or any other activity related to the course depending on its nature.
- Distribution of work between a team of students. Collaborative learning is a typical and valuable setting. Unfortunately, teachers are often frustrated about the lack of information about the actual work developed by each team member. Analysing the interactions and resources used by each of them can provide a better understanding of the team dynamics.
- In-class activity: despite being the most accessible scenario (as the teacher is actually present in the class and can observe the students directly), it is not as easy to monitor as it seems. Teachers are usually busy enough explaining the material, solving doubts and attending students' questions. In addition, mass classes make it hard to keep track of individual students' progress.

Previous work in the *dataTEL* domain emphasises the need of building a collection of learning data sets and issues related, from data gathering to normalising and sharing (Drachsler et al., 2010); as well as potential applications of such data sets (Verbert et al., 2011). In this work, we describe an unique experimental data set of learning activity events, the data gathering methodology and transformation process required for transforming it into an inter-operable format together with the corresponding supporting system, and a set of visualisations analysing such data aimed at supporting the teaching activity.

The rest of this document is organised as follows. Section 2 describes the scenario and Section 3 the problem tackled. In Section 4 the architecture of the system for capturing observations and creating the visualisations is described. A description of the collected data is included in Section 5 followed by a discussion of the obtained visualisations in Section 6. The paper concludes describing some conclusions and venues for future work.

2 Context Description

This work has been carried out in a typical Engineering course. In particular, the monitored data and discussed results refer to a Systems Architecture course, which

is part of the second year of the Telecommunication Engineering programme in the University Carlos III of Madrid (Spain). A total of 248 students enrolled in the course, which were divided into five sections. Four learning outcomes are included in the course. After taking the course, students must be able to (Pardo et al., 2010):

1. Design a software system using the C Programming Language containing non-trivial data structures, dynamic memory management, and using engineering techniques to translate a set of given high level constraints, derived from a hypothetical industrial setting, into a robust application,
2. Use proficiently the following industrial tools: a compiler, a debugger, a version controlled system, a cross compiler, and profiling tools to analyse memory behaviour.
3. Work effectively in a team to execute a project entailing the design of a software application on a mobile device, generate ideas collaboratively in a team to promote the exchange of information, organise the work in a team to optimise its performance, comply with the project requirements, and divide tasks effectively among the team members.
4. Learn autonomously, manage different information sources, generate and value concise information about the tasks accomplished, manage the time of personal work, and present effectively the results derived from the process.

The first two objectives (design and develop C applications, and use development tools) refer to knowledge, skills and competences related to the course domain. The two latter (team work and self-learning) emphasise methodological competences instead. All of them are considered within an active learning strategy. The course meets twice a week alternating lectures and laboratory sessions. Lectures cover the main theoretical concepts of the course. Laboratory sessions include hands-on sessions related to those theory concepts. The entire course adopts an active learning methodology. Each session has two sets of activities: one to be completed prior to the session, and the second to solve during the class. The final course score is obtained following a continuous evaluation scheme where questions, small exercises, and lab submissions are spreaded along the semester. The learning process combines different types of activities: readings, self-assessment tests, exercises, tool use, etc. All of them have in common that the student is always expected to play an active role, never being just a passive receptor of information.

3 Problem Statement

The problem tackled in this document is if a reliable observation mechanism capable of capturing a rich set of events occurring in a learning environment can be deployed, and the obtained data visualised intuitively so that certain aspects of that environment previously hidden can be shown to the teaching staff.

In an environment such as the one described in Section 2, the LMS plays only a marginal role. Most of the events occur while students are either working in the proposed procedures, with certain tools, and/or outside the classroom. This setting severely reduces the capacity of instructors to observe in detail the events

occurring while students work in the course activities. This scenario can benefit from the proposed approach where a set of visualisations are derived from the data collected while students use a virtual machine previously configured with all the tools required in the course. A better understanding of the overall class progress as well as individual evolution can help teachers to direct their efforts more efficiently. Early detection of problematic cases can also make a difference, allowing teachers to apply specific measures for diminishing failure rate.

4 A Layered Architecture for Learning Analysis

In this section, a generic layered architecture is described to collect events from different sources and combine them to obtain visualisations and support interventions. The considered layers are monitoring, pre-processing, data storage, analysis/visualisation, and intervention. Figure 1 depicts these layers and their connections.

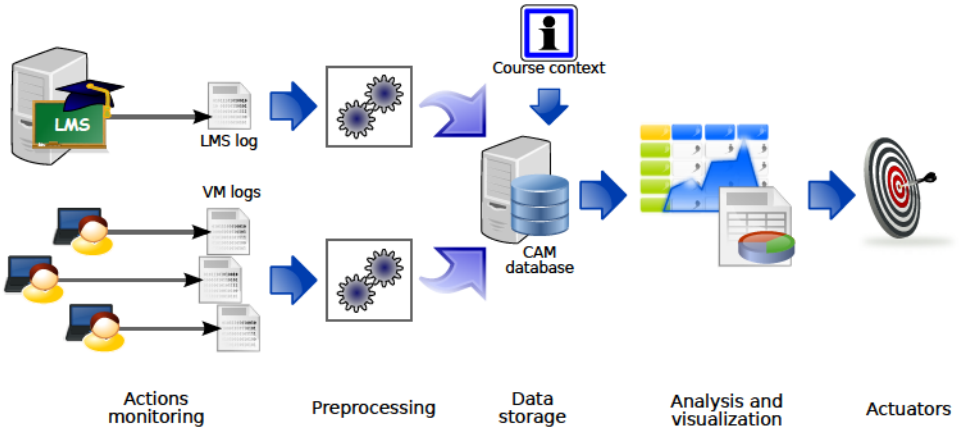

Figure 1 Layered Architecture for Learning Analytics

- Action monitoring. A network of sensors captures events related to student activity from different contexts (personal computer, LMS, etc.)
- Preprocessing. The raw data obtained from these sensors are transformed into a common representation model. In the proposed application the Contextualized Attention Metadata (CAM Schema, 2011) was proposed to promote the exchange of data among different institutions.
- Data storage. The processed data is stored in a relational database where additional information from the learning context can be added by other applications.
- Analysis and visualisation. The collected events are analysed and a set of visualisations are produced. The result of this analysis can also be used to detect interventions or recommendations automatically.

- Actuators. This layer includes different mechanisms to intervene in the learning process (namely actuators). For example, the detection of certain event patterns may automatically notify the corresponding instructor to review the situation.

This paper includes contributions in the first four layers of this architecture. It follows a description of each of them.

4.1 Distributed Network of Sensors

In the first layer of the architecture, the objective is to cover an area as wider as possible of the learning scenario. The proposed approach is to use the logs in the LMS and extend them with the events recorded while students use a fully configured virtual machine.

4.1.1 Server logs

Server logs provide a detailed account of the events occurring in a virtual community. However, in some cases, the granularity of these observations is too low. In other words, although a line in a server log denotes a simple event, the analysis and visualisation techniques proposed later in the document required these events to be grouped to identify higher level actions. For example, when a student enters in a forum, visualises a thread and posts an answer, a set of log entries are produced, when in fact, the only event that needs to be identified is that a student posted in a forum. Once the low level events were encoded, a grammar and a set of regular expressions were used to translate these sequences into higher single events. In this case, the virtual community of the course was hosted in dotLRN (Apache server). Processing any other LMS log would require adapting the grammar and regular expressions accordingly. Examples of resulting high level events are logging in, logging out, posting in a forum, viewing messages, replying to messages, etc.

4.1.2 Virtual Machine

Virtualisation is a paradigm by which a computer ("physical machine") is used to simulate the behaviour of a second computer ("virtual machine"). The use of these solutions in education has gained importance specially in those settings in which a collection of specific tools is required to work in a course. The main advantage of these virtual machines is that they can be previously pre-configured with a complete set of applications. Students may use their own personal computers and the only requirement is to install a virtualisation application.

This encapsulating technique opens the door to increase the type of activities that can be monitored. In the system designed for the described scenario, observation mechanisms were integrated in the virtual machine for monitoring the usage of the main tools required for the course (browser, programming tools, editor, etc.) (Romero Zaldívar et al., 2012). Students were provided the first day with such a fully configured virtual machine. As a result, a detailed account of each invocation of each tools for each student has been obtained. This approach is specially powerful in learning contexts with a high component of procedural activities. Due to the practical content of the course, this information can be used as the starting point

to deduct important facts about how students work in the previous activities or in tasks performed outside the supervision of an instructor. Students were informed about the type of events being recorded as well as the procedure to disable the tracking mechanism if they decide to do so, and their right to consult or delete any collected data in the future.

This configuration has several limitations. The first one is that students may choose to ignore this virtual machine and work in an environment in which no observation is possible. Students could not be forced to use the system, but their submissions required to work with the proposed tools. The alternative to not using the given appliance meant to configure the environment themselves. A small number of students selected this option. The second limitation is that even though the virtual machine is installed, students may still carry out activities in their physical machine, and therefore, they are not observed. Additionally, the student can choose to exercise their rights regarding data privacy disabling the monitoring system. Finally, in collaboration settings where students work together sharing the computer, all events in the session are associated to just one of the students of the team. This situation means that the observed events may provide a partial view. The visualisations discussed in Section 6 show that although the coverage might not be exhaustive, the information is still relevant for the teaching staff.

4.2 Data Storage

The distributed network of sensors (both the virtual machines and the server logs) generates raw data containing the information about the registered events. These files need to be processed in order to provide a coherent and integrated data representation for visualisations and analysis. The raw data need to be processed, complemented with contextual information and stored in an appropriate format to facilitate the analysis. In this work, the Contextualised Attention Metadata (CAM) format was used. The CAM schema allows modelling a user's handling of digital content across system boundaries (Schmitz et al., 2007).

CAM reflects the information about user's explicit actions rather than high level information about user's attention (which is expected to be derived from them). The reason to choose CAM to represent the collected information is because it was designed to be generic, simple and complete. By adopting a generic representation model, we are guaranteed that events are easily encoded from any potential new sensor. The model has the simple structure shown in Figure 2 (CAM Schema, 2011).

The central element of CAM is the `event`, which represents a user action, e.g. *User U compiles Program P with Application A generating the OuputMessage M.* Intrinsic characteristics of the event are stored as attributes: the `name` attribute represents the type of the event, e.g. open_document, compiler (compiling a program), visit_url (browsing a url), etc.; and the `dateTime` attribute represents the time-stamp of the event (e.g. $2011-10-0714:49:59$). Every `event` is connected to at least one `session`, which may include attributes such as `IP address`, `domain`, or a `session ID`.

Any additional elements related to an event (users involved, used application, etc.) are represented as `entities`. Thus, each event can be connected to several entities, and vice versa. The event-entity connection requires the role played by the entity in the event to be specified. `Entity` elements are defined by attributes

for `name` and, optionally, `MIMEtype`, `metadataId` and `metadataReference` (the two latter referring to metadata information specific to the entity itself, e.g. LOM metadata if available).

Being highly dependent on the scenario, context and analysis objectives, CAM does not specify a structure for contextual information. In this work, however, context information can play an important role in the analysis and thus needs to be included in the data model. In order to avoid modifications to the CAM schema and maintain compatibility, context metadata has been included as complementary tables in the relational structure.

The `metadatadescription` element represents the types of context information considered (in this work, the Program and the groups the students are structured into) and the `RelatedEntityMetadata` connects the entity elements with the corresponding contextual information.

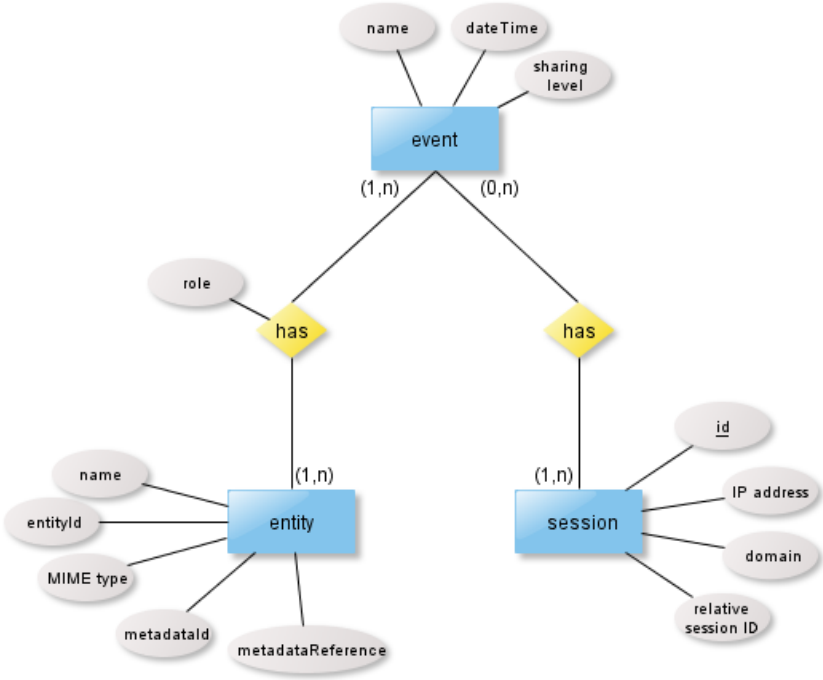

Figure 2 Contextualized Attention Metadata (CAM) Schema.

4.3 Data Visualisation and Analysis

Once the student actions and the LMS logs are preprocessed and stored as events in the CAM database, they can are analysed to produce useful visualisations information for the teaching staff. The collected data can be analysed and visualised from different perspectives, related to the diverse objectives and variables which the analysis is based on.

According to the classic model of interaction by Moore (1989), the learning process can be modelled as a system where three kinds of interactions stand out: learner-content, learner-tutor and learner-learner. It was later extended by Anderson and Garrison (1998) to consider interactions focused on the tutor and content too (see Figure 3). As technology-enhanced learning increases its importance in education, technology can be added as a fourth actor or at least a supporting layer.

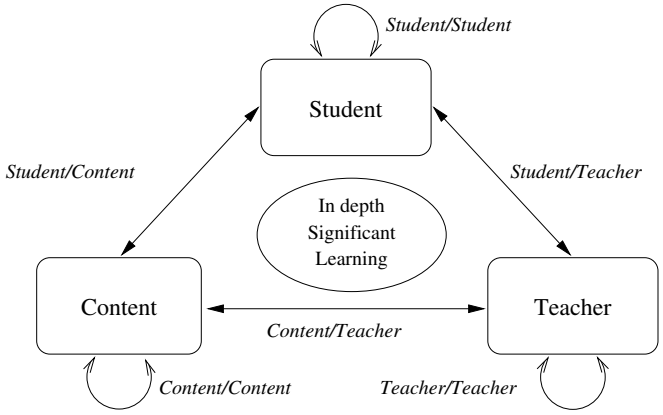

Figure 3 Moore's interaction model, extended by Anderson and Garrison.

This characterisation of the interactions occurring in a learning environment serves as a guide to obtain different visualisations. A first dimension to explore is the type of interaction. For example, events that are part of the interaction among students may show the social structure emerging from a learning community, or the level of collaboration among team members. Interaction between students and teachers may show useful patterns on the need of support through the course. Events recorded while students interact with the content can be used to visualise if learning skills such as information retrieval, planning and organisation, etc. are acquired. Finally, another dimension to explore in this layer is time. If certain intervention in a learning community is decided, its effect can be measured by the evolution in time of certain observations.

Visualisations can provide a first understanding of the underlying processes that are typically hidden to instructors. There are two main features defined for visualisations: interactivity and the type of graphics. Different kind of graphics allow to interpret the data from different perspectives. For example:

- Graph-based visualisations allow to represent the relations between the entities.
- Tree-based visualisations are useful for showing the distribution of data organised hierarchically in classes.
- Scatter plots provide detail information though they can be more difficult to understand due to data overload.
- Heatmaps provide a simple and clear summary of the distribution of items.

Providing complementary visualisations can foster deeper insight. Our goal in this work is to provide a set of visualisations that allows teachers to better understand the learning process.

In this layer of the architecture more complex post-processing and analysis can be applied. As these are beyond the scope of this paper, we simply mention some applications that are currently being explored. For example, resources from external sources can be discovered (and recommended to the students) analysing the browsing activity, as explained by Romero Zaldívar et al. (2011). Additionally, action patterns can be extracted from the student's activity and presented for self-reflection, as suggested by Scheffel et al. (2011). The visualisations proposed in this document are derived from the interactions among students, and the interaction between students and course resources.

5 Collected Datasets

The data set collected with the virtual machine as described in Section4.1.2 is summarised in Table 1.

Number of events recorded	101,972
Students enrolled in the course	248
Number of students with recorded events	172
Type of events	12

Table 1 Data set obtained with the Virtual Machine

The number of students for which some event was recorded amounts to 69% of the total course enrolment. A significant amount of students for which no event was recorded had dropped the course in the first week due to the special requirements imposed by the institution. The number of events obtained and the types of events show that events were observed with different levels of coverage depending on the use of the virtual machine. Also, data is relayed to a central server by invoking a version control system that students use to manage a folder shared among team members. Some of the events are recorded for one of the students while the rest may participate in the activity but no event is recorded. This is the reason why the dataset was used for instructor information, because the data about the group of students is still relevant. Table 2 shows the events in the virtual machine considered after discarding certain event types with a low number of observations.

In addition to the collected events, some contextual information was added to the data set. For example, the dates and times of the theory and lab sessions were included from the course schedule. This data was added to the model because this information was used to define the scope of the analysis. Another contextual information added is related to how students are grouped. The student entity is extended with an attribute to store the group name. This information allows visualisations to be restricted to the activity of one group instead of the whole section.

The second data set was derived from the activity registered in the learning management system and reflects only those events related to the course forum.

Event Type	Count	Percentage	Per student			
			Min.	Mean	Max.	SD
Command	29,452	32.85%	0	171.23	1003	205.34
Visit URL	27,076	30.20%	0	157.42	946	200.03
Compile	22,826	25.46%	0	132.71	870	154.20
Editor	7,449	8.31%	0	43.31	300	44.60
Profiler	2,466	2.75%	0	14.34	142	24.99
IDE	238	0.27%	0	1.38	42	3.87
Debugger	152	0.17%	0	0.88	26	2.79
Total	89,659	100.00%	0	521.27	2822	520.71

Table 2 Summary of the recorded events

The discussions that are part of any activity in the course are all done in face-to-face session, therefore, the forum in the virtual community is mainly used to ask questions to the instructors. Thus, the data set includes the messages posted in the forum, the authors, and the thread relationship. That is, if a message started a new discussion thread, or is an answer to a message in another thread. The content of the messages was not stored because the purpose of the data was to show the interaction patterns (as it will be discussed in Section 6).

The collected data convey information about the resources and tools used by the students as well as their sequences of activities during the learning process. Such data can be exploited for different applications. They can be directly explored and analysed, providing deeper insight on the learning process, as discussed in Section 6. Furthermore, they can be used as basis of recommender systems and for evaluating recommendation algorithms, as proposed by Verbert et al. (2011). Such an application is reported by Romero Zaldívar et al. (2011).

However, as remarked by Drachsler et al. (2010), "Privacy and legal protection rights are a challenging and rather important topic when talking about sharing of data sets. [...] Thus, before a data set can be shared, care must be taken to anonymise the data as much as possible. [...] It is important to keep in mind that with some dedicated investigation, even anonymised users and their items in the data set can often be matched to the data on the Web 2.0". This remains an unsolved problem, thus preventing us from making publicly available the described datasets, due to the restrictive constraints imposed by Spanish legislation (Spain, 1999).

6 Activity visualisations (with examples)

This section includes the portfolio of visualisations derived from the data gathered in the course described in Section 2. The target audience for this visualisations is the teaching staff. The diagrams can help them to better understand what is happening in their courses either by showing information unavailable until now, or by stating information explicitly about issues already known.

The visualisations presented in this section focus on different learning outcomes of the course, such as self learning, use of industrial tools, time management, information retrieval, collaboration, etc. Some of them integrate the information

from a set of students (typically all the class), while others focus on individual behaviour. However, it would be straightforward to filter the data in order to narrow the analysed set or to compare certain students (e.g. to see how good and bad students differ).

The analysis starts showing the data commonly available, i.e. the LMS logs. Then, the scope of the analysis is widened to consider the entire browsing activity of the students (of which the LMS logs reflect only a fraction). The analysis is further broadened in order to consider all kinds of learning tasks. Finally, the analysis focuses on specific learning activities, such as how the students are collaborating and interacting with the course tools or how a particular learning session in a lab develops. A detailed description of each visualisation follows.

6.1 Forum Activity

There are numerous tools already available that offer this type of visualisation. The students participating in the course under study were divided into sections, and each section had a separated forum. Figure 4 shows the graph of interactions derived from the LMS server logs for one of the forums in the course. Nodes represent students and teachers, and the edges relate students or teachers who participate in the same forum thread (writing messages). The more threads they collaborate in, the thicker the edge is drawn.

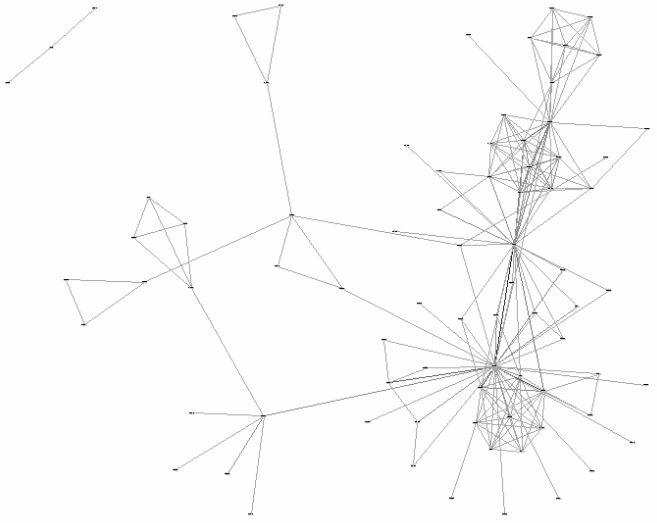

Figure 4 Forum interactions

6.2 Browsing activity

Although the LMS logs provide valuable insight of part of the student activity, numerous key learning activities happen outside its scope. The data obtained from

the virtual appliances provided to the students allow for a more in-depth analysis of these activities.

Figure 5 shows all the URLs that students typed while working in the course. URLs highlighted in green represent the course notes available in the LMS. A tree based visualisation has been applied in order to group the urls based on their origin (domain, server and path). This allows distinguishing local resources in the student's computer (file protocol) from those stored in remote web server. Also, URLs pointing to the LMS are highlighted in the figure (light blue).

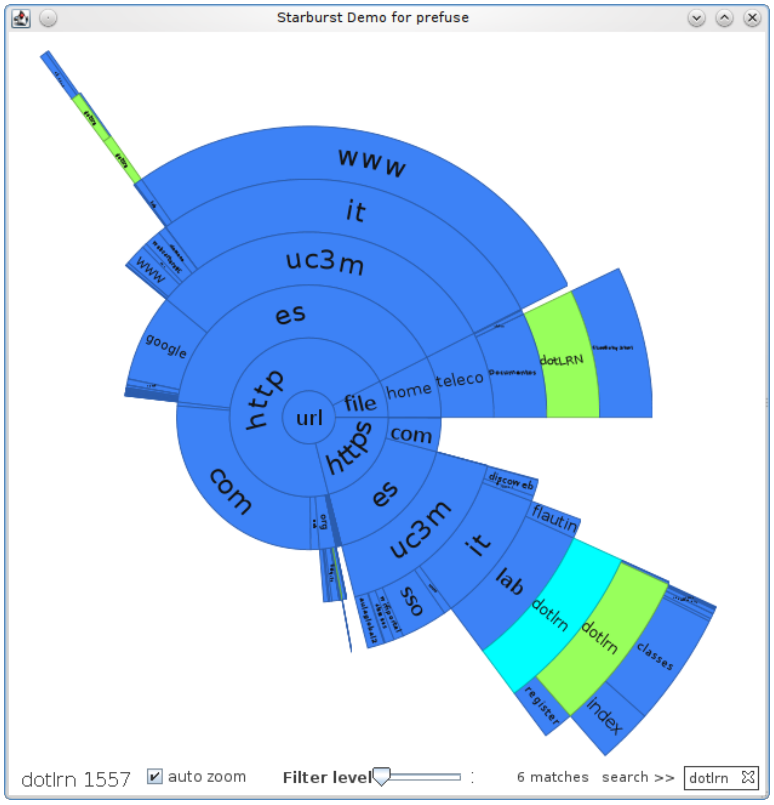

Figure 5 Browsing activity outside the LMS

This visualisation allows to see the distribution of the browsing activity while working in the course. Most of the browsing visits are outside the LMS. From a total of 8,669 unique pages, only the 28.51% point to the LMS, meaning that almost three out of four pages are outside the LMS. This is particularly relevant as it is not just other kind of activities (e.g. problem solving) the ones that are left out of the LMS, but even an important part of the browsing activity itself would be missed if the monitoring is limited to the system server. This definitely endorses the necessity of monitoring the student's activity in their own computers.

Sectors including the word "dotlrn" correspond to the activity in the LMS and have been highlighted in Figure 5. They include:

- The 'dotlrn' sector under 'file' (in the right, towards the middle of the figure) refers to the course materials stored locally in the student's virtual machine. The vast majority of the URLs browsed locally correspond to course materials.

- The 'dotlrn' sector under the 'https' section (down-right, highlighted in light blue) corresponds to visits to LMS resources. Just above half of such visits correspond to actual classes resources, as a significant part of them relate to the registration page or the course home page.

- Finally, the small, non-significant 'dotlrn' sector in the 'http' sector (up-left corner) groups the accesses to non-secure pages referring to the LMS.

The rest of the visits correspond to other resources external to the course. An important fraction corresponds to the departmental web server or other institutional resources. Nevertheless, there is still a significant fraction corresponding to totally unrelated servers, including social networks (e.g. Tuenti), search engines (e.g. google.es), information websites (e.g. Wikipedia), technical forums, etc.

The visualisation of the student browsing activity can help the teaching staff see the type of resources uses for the activities as well as the preferred technological contexts for the students. This information may serve to deploy more appropriate remedial actions.

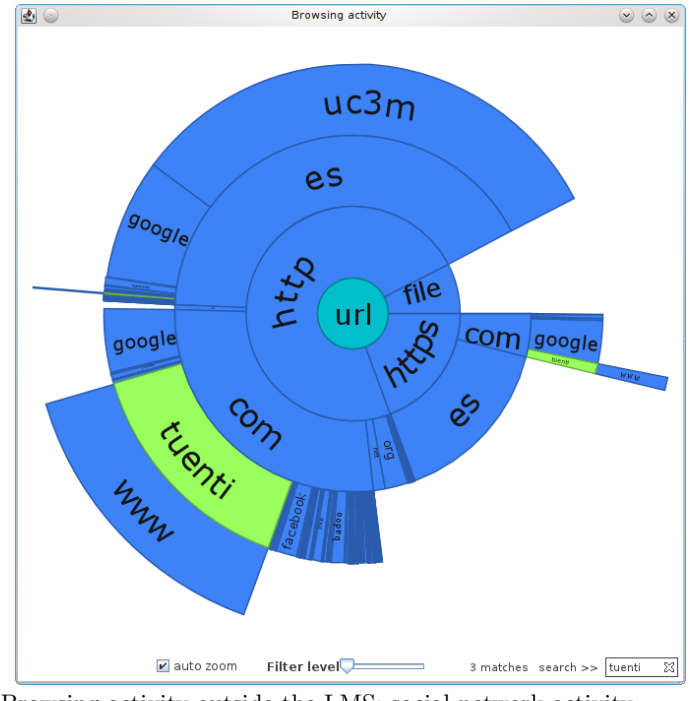

Figure 6 Browsing activity outside the LMS: social network activity

Site	Category	#Hits	%Hits
University domain	Institutional	7756	50.02 %
Tuenti	Social network	2363	15.24 %
Google	Search, docs and others	2204	14.21 %
Facebook	Social network	202	1.43 %
Twitter	Micro-blogging	13	0.08 %
Mail (including the university one)	Mail	540	3.57 %

Table 3 Hits for most typical services.

6.3 Use of Social Networks

Digging into the interactive visualisation of browsing activity, Figure 6 highlights the relevance of social networks for this cohort of students. Apart from the university website, the most visited website is Tuenti (tuenti.com), a Spanish social network for people around twenty years of age, closely followed by Google (google.com or google.es) and far beyond other sites like Facebook. The data highlights the importance of social networks and also shows which ones are they using. If the teaching staff decides to extend the course with a page in a social network, Tuenti would then seem the most appropriate. Furthermore, micro-blogging applications such as "twitter" has a very low presence. Teaching staff pondered for some time the need to deploy a course channel in "Twitter" to notify of events related to the course, but it was ultimately discarded after seeing these data. Also derived from the figure, it is interesting to see the decreasing trend in the use of e-mail, closely related to the social networks boom. Table 3 details the number of visits for some well-known services.

The importance of this kind of analysis of the students' technological background is due to its dynamic nature. First, the evolution of technology itself makes necessary a continuous reevaluation of the context. What applies this year will probably change for the next one. Second, it depends greatly on the profile of the students themselves. While this cohort of students are not using Twitter, it can still be a popular tool for future editions of the course or other studies.

6.4 Activity outside the LMS

Conclusions from the analysis of the students' browsing activity become even more remarkable when comparing to the global distribution of events based on their type, represented in Figure 7.

Events related to browsing activity are 32.07%, around 15,500 out of a total of approximately 48,000 events that occurred in the first half of the course. Out of that, 1,557 correspond to hits related to the LMS. This sums up a 10% of the total number of hits and just a 3.21% of the total actions, meaning that 96.79% of the events correspond to actions outside the LMS.

The starburst visualisation in Figure 7 can be explored interactively, expanding and collapsing the sections corresponding to the events organised hierarchically, increasing and decreasing the depth, and searching for specific terms. Although the snapshot shown in Figure 7 represents just a summary of the distribution of the types of events, some interesting insights emerge. For example, compiler and bash

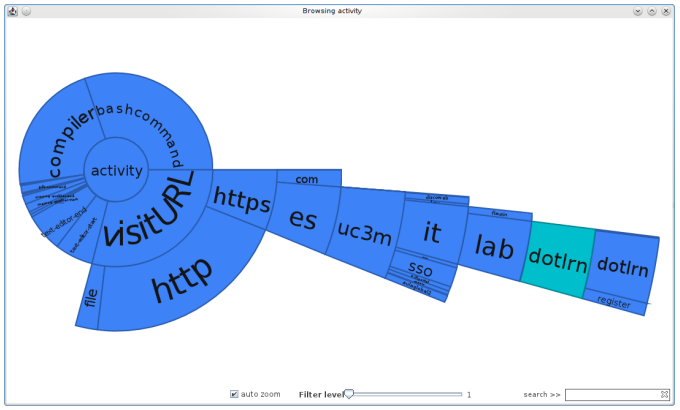

Figure 7 Students' work outside the LMS

commands conform the biggest classes together with the browsing events, further away from the rest of the event types. This seems reasonable, considering the course contents.

However, some classes appear shockingly reduced. Edition events monitoring is limited to starting and closing the editor, which explains a class size smaller than what could expected. Nevertheless, debugging events are clearly missing in the picture: Only 53 debugger sessions (corresponding to 53 debugger-start and debugger-end events) with 594 events corresponding to gdb commands, compared to a total of 10,412 compiler events (less than 6 %). As it cannot be explained due to the lack of monitoring, students are actually using the debugger scarcely, despite the emphasis put on this tool during the course. This is confirmed by the teaching staff

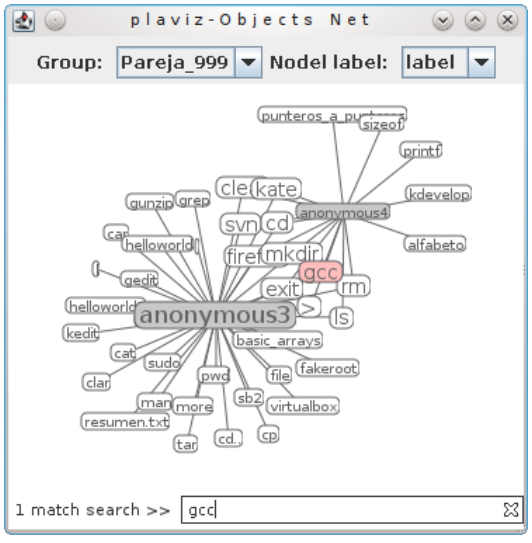

Figure 8 Resources used by a team of two students

6.5 Single team analysis

Figure 8 visualises the tools used by a team of two students during the first sessions. Students work in pairs in the lab, which means that usually all events in a session are associated to just one of the students of the team, as explained in Section 4.1.2. Thus, several sessions have been selected instead of just one in order to get a fairer grasp of the distribution of the work. The nodes represent the students and the resources they interact with. The size of the nodes is proportional to the number of events they appear in. This allows to easily compare students' activity as well as tools use (or lack of).

6.6 Activity during a lab session

Figure 9 provides a global representation of the events during a lab session, grouped in rows by student. The purpose of this visualisation is to help instructors analyse how effective a laboratory session has been for the different students. The horizontal axis represents time, more precisely the 120 minutes (90 minute session with 15 minutes before and 15 after class time) in which the session took place. The vertical axis is divided into five rows, each of them containing the events recorded for a specific student. Each point in the row corresponds with an event that took place at the specified time. All the events are shown without distinguishing their types. The idea of this representation is to provide an intuitive visualisation of the overall activity of the student in the session.

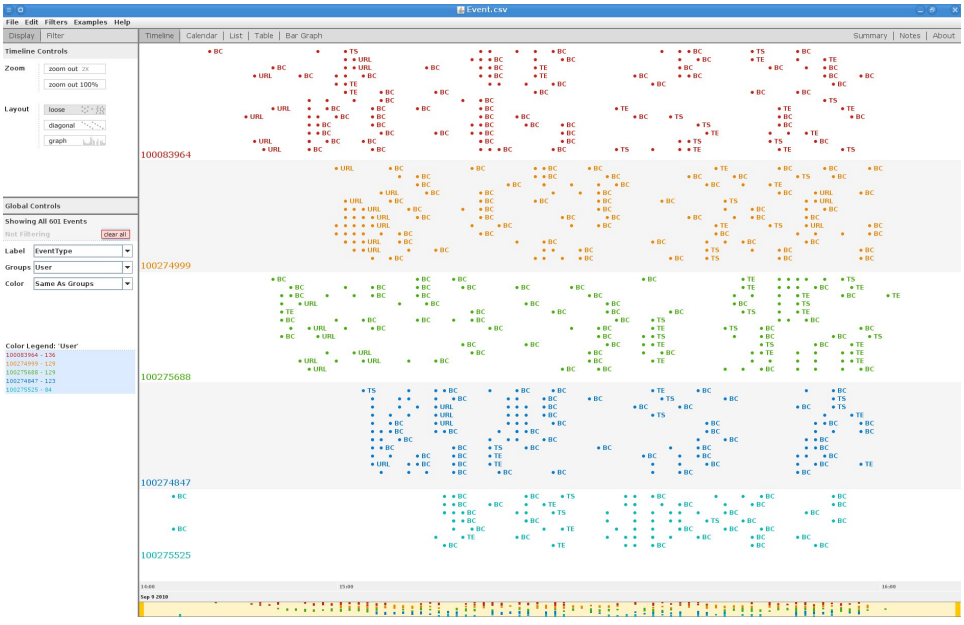

Figure 9 In-class activity per student

There are various aspects that are clearly shown. Ideally, students should start their activity immediately after a brief explanation by the teacher at the beginning

of the class and work continuously along the session. For example, the student at the top of the graph has worked during the whole time when the session was taking place at a sustained pace. The density of events all throughout the session seem to be balanced. A different situation occurs with the student represented at the bottom. The plot clearly shows that the student did not engage in the class until it was well under way. Additionally, the level of participation is lower than the rest of classmates. Second and fourth students from the top also have a less than ideal activity during the session, taking longer than expected to begin working. Although instructors may perceive this information during the class, this type of visualisation gives them a formal observation point that can be used to prescribe immediate remediation actions. Furthermore, this information can be redirected to the student, to promote self-reflection and encourage self-organisation and self-learning.

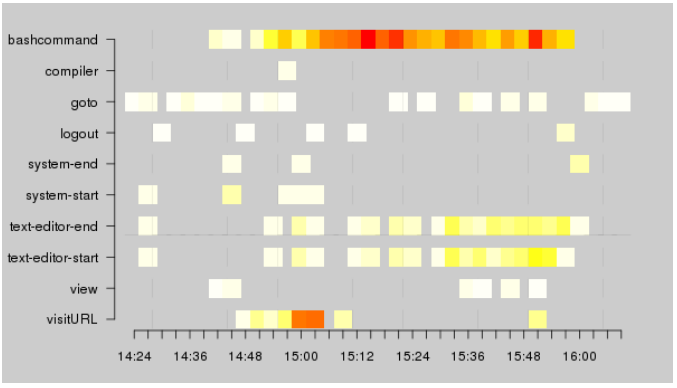

Figure 10 In-class activity heatmap per type of event

6.7 Activity of a student during a lab session

Figure 10 is another provided visualisation that further develops the information in Figure 9, containing a more detailed report of the activity for a single student. In this case, the events are clearly segregated by type thus offering a finer granularity for the observation. Each event type is shown as a row in the graph. Note that they have been renamed to terms more familiar to the teacher. The colour represents the number of events of such type that were recorded in that time unit. The data is normalised with respect to the entire group. That is, the red colour stands for the maximum account of all the events. As in the previous visualisation, the horizontal axis is the time.

As it can be seen, most of the activity occurs with events of the type corresponding to the top row (*bash command*). This is consistent with the type of activity requested in that session, devoted to the bash shell, which also explains the lack of compiling events (*compiler*). The third row (*goto*) are visits to the virtual community stored in the LMS, while *view* events refer to visualising a forum message. Note that log in the LMS is not shown as it is assumed to happen together with the first LMS event, and only *logout* from the LMS events are shown explicitly. Launching and closing the virtual machine are registered as *system-start*

and *system-end* events, respectively. Similarly, starting and closing the text editor are registered as *text-editor-start* and *text-editor-end*, respectively. The bottom row (*visitURL*) is the number of URLs visited by the student outside of the LMS. This row shows that there is a high activity of these events shortly after the beginning of the session and then most of the activity due to the use of tools (an editor) and the execution of various commands.

7 Conclusions and Future Work

In this paper, a data collection technique and a set of visualisations have been presented, together with the experimental results derived from their deployment in a real course. The objective was to extend the observation and analysis capabilities beyond what happens in a Learning Management System. Most learning experiences require the students to work with additional tools and in settings that are detached from the LMS. The proposed monitoring scheme combines the use of a virtual machine fully configured and offered to the students to use for the course activities, with the use of the events recorded at by the LMS, to provide a more complete coverage of the learning process.

These techniques have been deployed in a face-to-face second year engineering course. Two data sets were obtained: one from the use of virtual machines, and a second one from the LMS. The described visualisations confirm the relevance of events occurring outside of the LMS. The strategy followed by students when accessing the course material, the intensity with which they solve the proposed exercises in a laboratory, the type of tools that are being used, are all now visible with the new data. Both teachers and students are expected to benefit from the information provided, which is meant to support self-reflection, awareness and decision making.

With these visualisation, a variety of future avenues are open for research. Instructors and students alike can react to the mere presence of this information. Technology may support this reactions by providing a set of "actuators" that are either automatically enabled, or offered to the stakeholders so that a "person in the loop" is maintained. A parallel research line works on the development of complementary analysis applying multidimensional analysis (OLAP) and data mining techniques.

Acknowledgements

Work partially funded by the projects: *Adaptation of learning scenarios in the .LRN platform based on Contextualized Attention Metadata (CAM)* (DE2009-0051), *Learn3* ("Plan Nacional de I+D+I" TIN2008-05163/TSI), *EEE* ("Plan Nacional de I+D+I" TIN 2011-28308-C03-01), and *Emadrid: Investigación y desarrollo de tecnologías para el e-learning en la Comunidad de Madrid* (S2009/TIC-1650).

References

Anderson, T. D. and Garrison, D. R. (1998). Learning in a networked world: New roles and responsibilties. In Gibson, C., editor, *Distance Learners in Higher Education*, chapter 6, pages 97–112. Atwood Publishing.

Arnold, K. (2010). Signals: Applying Academic Analytics. *EDUCAUSE Quarterly*, 33(1):10.

Auinger, A., Ebner, M., Nedbal, D., and Holzinger, A. (2009). Mixing content and endless collaboration–MashUps: Towards future personal learning environments. *Universal Access in Human-Computer Interaction. Applications and Services*, pages 14–23.

CAM Schema (2011). https://sites.google.com/site/camschema/.

Chatti, M., Jarke, M., and Frosch-Wilke, D. (2007). The future of e-learning: a shift to knowledge networking and social software. *International journal of knowledge and learning*, 3(4):404–420.

Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., Lindstaedt, S., Stern, H., Friedrich, M., and et al. (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1(2):2849–2858.

Ertugrul, N. (2000). Towards virtual laboratories: A survey of LabVIEW-Based teaching/learning tools and future trends. *International Journal of Engineering Education*, 16(3):171–180.

Goldstein, P. and Katz, R. (2005). Academic analytics: the uses of management information and technology in higher education. Technical Report December, EDUCAUSE Center for Applied Research.

Govaerts, S., Verbert, K., Klerkx, J., and Duval, E. (2010). Visualizing activities for self-reflection and awareness. In *The 9th International Conference on Web-based Learning, ICWL 2010*, Lecture Notes on Computer Science. Springer.

Janssen, J., Erkens, G., Kanselaar, G., and Jaspers, J. (2007). Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Computers & Education*, 49(4):1037–1065.

Lin, F., Hsieh, L., and Chuang, F. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2):481–495.

Long, P. and Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *Educause Review*, 48(5):31–40.

Mazza, R. and Dimitrova, V. (2004). Visualising student tracking data to support instructors in web-based distance education. In *Proc. of the 13th Int. World Wide Web Conf. (WWW 2004) - Educational Track*, pages 154–161. ACM.

Moore, M. G. (1989). Three types of interaction. *The American Journal of Distance Education*, 3(2):1–6.

Pardo, A., Estévez-Ayres, I., Basanta-Val, P., and Fuentes-Lorenzo, D. (2010). Programación en C con aprendizaje activo, evaluación continua y trabajo en equipo: caso de estudio. In *Jornadas de Enseñanza Universitaria de la Informática*, pages 231–238.

Romero Zaldívar, V. A., Crespo García, R. M., Burgos, D., Delgado Kloos, C., and Pardo, A. (2011). Automatic Discovery of Complementary Learning Resources. In Delgado Kloos, C., Gillet, D., García Crespo, R. M., Wild, F., and Wolpers, M., editors, *Proceedings of the European Conference on Technology Enhanced Learning. LNCS Vol. 6964*, pages 327–340. Springer.

Romero Zaldívar, V. A., Pardo, A., Burgos, D., and Delgado Kloos, C. (2012). Monitoring Student Progress Using Virtual Appliances : A Proof of Concept. *Computers & Education*, To appear:1–18.

Scheffel, M., Niemann, K., Pardo, A., Leony Arreaga, D., Friedrich, M., Schmidt, K., Wolpers, M., and Delgado Kloos, C. (2011). Usage Pattern Recognition in Student Activities. In Delgado Kloos, C., Gillet, D., García Crespo, R. M., Wild, F., and Wolpers, M., editors, *Proceedings of the European Conference on Technology Enhanced Learning. LNCS Vol. 6964*, pages 341–355. Springer.

Schmitz, H., Scheffel, M., Friedrich, M., Jahn, M., Niemann, K., and Wolpers, M. (2009). Camera for PLE. In *EC-TEL*, volume 5794 of *LNCS*.

Schmitz, H.-C., Wolpers, M., Kirschenmann, U., and Niemann, K. (2007). Contextualized Attention Metadata. In *Human Attention in Digital Environments*, volume 13, chapter 8. Cambridge University Press.

Spain (1999). Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. *Boletín Oficial del Estado*, núm. 298 de 14/12/1999: págs. 43088–43099. BOE-A-1999-23750.

Tanes, Z., Arnold, K. E., King, A. S., and Remnet, M. A. (2011). Using Signals for appropriate feedback: Perceptions and practices. *Computers & Education*, 57(4):2414–2422.

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikar, i. R., and Duval, E. (2011). Dataset-driven Research for Improving Recommender Systems for Learning. In *1st International Conference Learning Analytics and Knowledge (LAK 2011)*, New York. ACM Press.

Waycott, J., Bennett, S., Kennedy, G., Dalgarno, B., and Gray, K. (2010). Digital divides? Student and staff perceptions of information and communication technologies. *Computers & Education*, 54(4):1202–1211.

Wolpers, M., Najjar, J., Verbert, K., and Duval, E. (2007). Tracking actual usage: the attention metadata approach. *Educational Technology & Society*, 10(3).

Woolf, B. (2008). *Building intelligent interactive tutors*. Morgan Kaufmann.

Zhang, H., Almeroth, K., Knight, A., Bulger, M., and Mayer, R. (2007). Moodog: Tracking students' online learning activities. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*.