Chapter 15
Data Mining and Social Network Analysis in the Educational Field: An Application for Non-Expert Users

Diego García-Saiz, Camilo Palazuelos and Marta Zorrilla

Abstract With the increasing popularity of social networking services like Facebook, social network analysis (SNA) has emerged again. Undoubtedly, there is an inherent social network in any learning context, where teachers, learners, and learning resources behave as main actors, among which different relationships can be defined, e.g., "participate in" among blogs, students, and learners. From their analysis, information about group cohesion, participation in activities, and connections among subjects can be obtained. At the same time, it is well-known the need of tools that help instructors, in particular those involved in distance education, to discover their students' behavior profile, models about how they participate in collaborative activities or likely the most important, to know the performance and dropout pattern with the aim of improving the teaching–learning process. Therefore, the goal of this chapter is to describe our E-learning Web Mining tool and the new services that it provides, supported by the use of SNA and classification techniques.

Keywords Data mining · Educational data mining · Social network analysis · Learning analytics

Abbreviations

API	Application programming interface
DM	Data mining
EDM	Educational data mining

D. García-Saiz · C. Palazuelos · M. Zorrilla (✉)
Department of Mathematics, Statistics, and Computer Science, University of Cantabria,
Avenida de los Castros s/n 39005 Santander, Spain
e-mail: marta.zorrilla@unican.es

D. García-Saiz
e-mail: diego.garcia@unican.es

C. Palazuelos
e-mail: camilo.palazuelos@unican.es

A. Peña-Ayala (ed.), *Educational Data Mining*,
Studies in Computational Intelligence 524, DOI: 10.1007/978-3-319-02738-8_15,

ElWM	E-learning web miner
KDD	Knowledge discovery in databases
LA	Learning analytics
LMS	Learning management system
MOOC	Massive open online course
SNA	Social network analysis
SOA	Service-oriented architecture
SOAP	Simple object access protocol
UC	University of Cantabria
WSDL	Web services description language
WS	Web service
XML	eXtended Markup Language

15.1 Introduction

Since the late 1990s, the use of computer-based technologies has drastically changed learning and teaching processes in all academic levels, from elementary school to university. Nowadays, it is very frequent that teachers include in their subjects activities that require the use of Web 2.0 technologies in order to develop contents and social and communication skills.

Collaborative activities, e.g., content search [1, 2], collaborative writing [3], and discussion forums [4], appear in many curricula independently of the educational field and level of the studies. Other tools frequently used, regardless of whether teaching is face-to-face or virtual, are the learning management systems (LMS), e.g., Moodle [5], Blackboard [6], or Shakai [7], which offer different modules, e.g., blogs, wikis, or forums, to develop collaborative activities that enable students to adapt to new environments and work in heterogeneous teams. This new scenario, where the degree of interaction among different actors, e.g., learners, educators, and resources, is very high, poses new situations and needs to instructors.

They need to know the students' level of cohesion, their degree of participation in forums, the identification of the most influential ones, which students help their classmates, and so on. This information might be helpful for teachers to organize team-works with different social profiles, grade the activities performed by their students according to their contribution, or spread news or relevant explanations through the most influential students. Especially, the analysis of social interaction might help teachers to better understand their students' social behavior and, as a consequence, assist them to improve their skills, as well as their results in the subjects involved.

Hence it is necessary to develop applications that help teachers to extract and analyze interaction data produced in the different teaching activities and their impact on student performance. This application must fulfill some requirements with the aim of being useful for non-expert users in the learning analytics field.

The 2013 Horizon Report [8] describes learning analytics (LA) as the "Field associated with deciphering trends and patterns from educational big data, or huge sets of student-related data, to further the advancement of a personalized, supportive system of higher education." This is a very wide field in which different techniques and tools are used by educators for gaining insights into student interaction with online texts and courseware and, consequently, being able to take actions to improve the teaching process.

This field comprises, among others, techniques from the educational data mining (EDM) field, which deals with the development and application of computational methods to detect patterns in large collections of educational data. Its main goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and gain insights into and explain educational phenomena [9]. Techniques from the SNA field are also employed in the academic context since the analysis of the structure and composition of relationships in the network provides useful information on the cohesion of individuals, their level of participation, or which individuals are the most active or influential.

Regrettably, both research fields use techniques and algorithms that make them unsuitable for people outside the fields of mathematics or computer science. Therefore, these algorithms and the associated processes for their execution must be wrapped in such a way that the end users should only worry about interpreting the results.

Furthermore, another key feature of this analytic tool is that it is independent from specific web applications or any other resource available that can be used in a collaborative activity. In such a way, it can be easily extended and enhanced as well as be used in different scenarios.

For example, LMS or massively open online course (MOOC) platforms are tools in which its inclusion would be very valuable since educators can design all the teaching process inside them. Furthermore, these platforms generally collect users' interaction—when students connect, how often, or when they write a post or perform a test—in databases that make it easier its utilization for the analysis processes. Although these platforms offer some monitoring tools, these are limited enough and, as stated by Macfadyen et al. [10], instructors in the new world of education are in need of new tools and strategies that will allow them to quickly identify students at risk and devise ways of supporting their learning.

In order to contribute to fill this gap, this chapter describes our enhanced version of E-learning Web Miner (ElWM) [11] and, more specifically, the new options that allow educators to gain insights into interaction, social behavior, and performance. Among other services, ElWM offers the generation of predictive models of students' performance based on classification techniques (in particular, decision trees), descriptive models for the characterization of learners from a social perspective based on SNA, graphs showing students' cohesion and those ones who are the most influential by using graph mining techniques, specifically FRINGE [12]. Likewise, we show its usefulness and simplicity through the analysis of different virtual courses and collaborative activities developed in both Moodle and Blackboard platforms at the University of Cantabria (UC).

This chapter is organized as follows. In Sect. 15.2, we provide a brief introduction to the context of our work, the theoretical foundation to understand our work and relate works published in the field of SNA applied to the educational context. We also cite the most relevant works focused on the generation of models of students' performance and dropout, and briefly describe other tools similar to our ElWM and discuss the main differences found. Section 15.3 describes the purpose of our tool, its architecture and the new services provided. Section 15.4 presents and discusses several case studies with the aim of showing the usefulness, simplicity, and added value that ElWM provides to the educational context. Finally, we summarize the contents of this chapter and discuss our future work.

15.2 Background and Related Work

EDM is an interdisciplinary area in which methods and techniques from computer science, education, and statistics are combined. LA is another field very related to EDM and with which it shares some goals and interests. However, LA integrates a broader array of academic disciplines, e.g., computer science, information science, learning sciences, psychology, sociology, and statistics.

Although there is no hard and fast distinction between these two fields [13], EDM focuses on developing methods and applying techniques from statistics, machine learning, and data mining (DM) to analyze data collected during teaching and learning with the aim of answering questions related to the educational practice, e.g., "What sequence of topics is the most effective for a specific student?" "What student actions are associated with more learning?" "What features of an online learning environment lead to better learning?" or "What will predict student success?" To accomplish its goals, EDM mainly uses methods based on prediction, classification, clustering, and association.

On the other hand, LA emphasizes measurement and data collection as activities that institutions need to undertake and understand, and focuses on the analysis and reporting of data. That means that LA, unlike EDM, does not develop new algorithms for data analysis but addresses the application of known methods and models in order to answer important questions that affect student learning and organizational learning systems.

Hence LA does not only focus on student performance, but it is also used to assess curricula, programs, and institutions. LA uses techniques related to concept, discourse, influence, and sentiment analyses, as well as sense-making models, SNA, statistics, and visualization [13].

Thus, our work fits in both fields since our tool uses exploratory and analytical techniques, as well as provides patterns generated with DM techniques. Specifically, we show the utility of SNA and its contribution to predict students' performance.

In this section, we provide some background on classification and SNA techniques and relate some of the most important works that apply DM techniques to discover prediction models of students' performance and dropout and SNA techniques to understand the interactions of students in e-learning courses. Furthermore, we include a section about tools developed for these purposes.

15.2.1 Social Network Analysis

SNA is the methodical study of the relationships present in connected actors from a social point of view. SNA represents both actors and relationships in terms of network theory, depicting them as a graph or network, where each node corresponds to an individual actor within the network, e.g., a person or an organization, and each link symbolizes some form of social interaction between two of those actors, e.g., friendship or kinship.

Although social networks have been studied for decades [14, 15], the recent emergence of social networking services like Facebook or Twitter has been the cause of the unprecedented popularity that this field of study has now. Since then, an extraordinary variety of SNA techniques has been developed, allowing researchers to model different types of interactions, e.g., movie actors [16] or sexual contact networks [17], and giving solution to very diverse problems, e.g., detection of criminal and terrorist patterns [18] or identification of important actors in social networks [19].

In order to estimate the prominence of a node in a social network, many centrality measures have been proposed. The research devoted to the concept of centrality addresses the question "Which are the most important nodes in a social network?" Although there are many possible definitions of importance, prominent nodes are supposed to be those that are extensively connected to other nodes. Generally, in social networks, people with extensive contacts are considered more influential than those with comparatively fewer contacts. Perhaps, the most simple centrality measure is the *degree* of a node, which is the number of links connected to it, without taking the direction of the links into consideration. If we consider that direction of the links, a node has both *indegree* and *outdegree*, which are the number of incoming and outgoing links attached to it, respectively. There are more complex centrality measures, such as the *betweenness* [20] of a node, which is equal to the number of shortest paths from all nodes to all others that pass through such a node.

Mathematically, let g_n^{pq} [1] be 1 if node n lies on the shortest path from p to q and 0 otherwise. Then the betweenness centrality of a node is given by (15.1).

$$b_n = \sum_{pq} g_n^{pq} \tag{15.1}$$

Authorities and *hubs* [20] are also two examples of more complex centrality measures; a node is an authority if its incoming links connect it to nodes that have a large number of outgoing links, whereas a node is a hub if its outgoing links connect it to nodes that have a large number of incoming links. Mathematically [see Eq. (15.2)], the authority centrality of a node is defined to be proportional to the sum of the hub centralities of the nodes that point to it, where α is a constant and A_{nm} is an element of the adjacency matrix of the network.

$$a_n = \alpha \sum_m A_{nm} h_n \tag{15.2}$$

Similarly, the hub centrality of a node [see Eq. (15.3)] is proportional to the sum of the authority centralities of the nodes it points to, where β is another constant.

$$h_n = \beta \sum_m A_{mn} a_m \tag{15.3}$$

From a node-level point of view, centrality measures constitute a very useful tool for the inference of the importance of nodes within a network. Due to their own nature, some of them, e.g., betweenness, cannot be trivially calculated, so that network-level metrics—which can be computed more easily and provide helpful information by considering the network as a whole—can be used for complementing the aforementioned centrality measures.

One of these network-level metrics is the density of the network, which measures the number of links within the network compared to the maximum possible number of links. The diameter of the network is also a useful network-level metric; it is defined as the largest number of nodes that must be traversed in order to travel from one node to another. Other meaningful network-level metric is the number of connected components of the network, i.e., the number of subnetworks in which any two nodes are connected to each other by paths without taking the direction of their links into consideration. Finally, the last metric to be mentioned is reciprocity; it occurs when the existence of a link from one node to another triggers the creation of the reverse link.

SNA techniques do not just concentrate on social networks, but also focus on other fields, such as marketing (customer and supplier networks) or public safety. Another field of application is education, although it is not deeply explored yet. There are some case studies in the literature, for instance: Brewe et al. [21] used a multiple regression analysis of the Bonacich centrality to evaluate the factors that influence participation in learning communities, e.g., students' age or gender. Crespo and Antunes [22] proposed a strategy to quantify the global contribution of each student in a team-work through adaptations of the PageRank algorithm.

Cuéllar et al. [23] proposed a method for the formulation and interpretation of learning management platforms as social networks with the aim of making further studies about the social structure among learners, teachers, and learning resources, and discovering useful relationships to improve the learning process. Rabbany et al. [24] built Meerkat-ED, a tool designed to assess the students' participation in asynchronous discussion forums of online courses.

SNA has also been used for extracting relevant information that can be used in EDM tasks. For example, Dawson et al. used SNA to monitor the learners' creative capacity [25], to detect and encourage students at risk [26]. Obsivac et al. [27] used several centrality measures for the prediction of dropouts. A similar work was performed by Palazuelos et al. [28] but, in this case, the goal was to evaluate whether SNA attributes are useful to build more accurate classification models.

15.2.2 *Classification Applied to the Educational Context: Students' Performance and Dropout*

DM is the process of automatically discovering useful information in large data repositories. DM is an integral part of Knowledge Discovery in Databases (KDD) [29] which is the overall process of converting raw data into useful information. This process comprises several steps: data selection and preparation, technique selection, testing and result evaluation. This process is not trivial; whenever a very accurate model is needed, one must turn to an expert in DM. Our tool will not achieve the most accurate model, but a reasonably good one at a very low cost that allows non-expert users to take advantage of using data mining towards its democratization.

DM techniques are generally divided into two major categories: predictive and descriptive tasks, being the former the one that arouses most interest. The prediction task consists of extracting relevant features from labeled training data to build a model that discriminates between classes to classify unlabeled observed objects. Prediction methods are, in turn, divided into classification and regression techniques. The former are used when the predicted variable is a categorical value and the latter, when the predicted variable is a continuous value or a probability density function. In EDM, classification techniques are more used, particularly, to forecast student performance [30–33] and detect anomalous students' behaviors, such as dropout [27, 34, 35].

Classification is the task of learning a target function f that maps each attribute set x to one of the predefined class labels y. There are different techniques to identify the model that best fits the relationship between the attribute set and the class label of the input data. Some examples are: decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naïve Bayes classifiers. The decision of which one to use depends on the DM goals that decision makers pursue. In the EDM field, we find some works in the scientific literature in which the authors try to establish which algorithms are the best to predict the students' performance and dropout.

Dekker et al. [34] presented a case study to predict student dropout in which demonstrated the effectiveness of several classification techniques and the cost-sensitive learning approach on several datasets with about 500 instances with numerical and nominal attributes that corresponded to pre-university and university characteristics.

Their experimental results showed that rather simple and intuitive classifiers, referring to decision trees, give a useful result with accuracies between 75 and 80 %. Kotsiantis et al. [35] also compared six classification algorithms to predict dropouts. Their comparison showed the Naïve Bayes algorithm [36] was the most appropriate.

Regarding predicting performance, a similar study was performed by Hämäläinen et al. [37]. In this paper, the authors compared five classification methods: multiple linear regression and support vector machine techniques to predict the numerical mark and three variations of the naïve Bayes classifier to predict the categorical qualification (pass/fail), concluding that all methods achieve about the same accuracy. Zafra et al. [33] compared several algorithms based on multi-instance learning to predict student's performance using data from a Moodle platform and the method that they proposed, G3P-MI [38], got the best results.

Finally, two of the authors of this work [38] compared five classification algorithms in order to determine which was the most suitable for educational datasets from e-learning platforms. Their experimentation concluded that there was no algorithm that achieved significantly better accuracy. When the dataset was very small (less than 100 instances) and had numeric attributes, naïve Bayes performed adequately; on the other hand, when the data set was bigger, BayesNet TAN [39] was a better alternative. However, J48 (implementation of the C4.5 [40] algorithm in Weka [41]) was suitable for datasets with more instances and/or with the presence of nominal attributes, being also the most interpretable. Thus, J48 was the algorithm chosen to be wrapped in ElWM. The process of building the classifier comprises two steps: training and test. When the dataset size is reduced, a good practice is to evaluate the performance of the classifier by means of cross-validation. Finally, the model must be evaluated. This task is based on the number of test records correctly and incorrectly predicted by the model. The most frequently used metrics are *accuracy*, *specificity* and *sensitivity*. Accuracy measures the number of correct predictions made by the model divided by the total of predictions. Sensitivity (true positive rate) measures the proportion of actual positives which are correctly identified as such. And specificity (true negative rate) measures the proportion of negatives that are correctly identified.

15.2.3 Data Mining Tools for Non-expert Users

The EDM field provides a large quantity of techniques and tools to further understand students and the settings which they learn in. In the last decade, many works have been carried out, as can be read in this survey [42]. Regrettably, most of these methods and tools are not directly used by non-experts in data mining, e.g., teachers. To fill this gap, our research group developed ElWM [43].

It aims to help instructors involved in distance education to discover their students' behavior profiles and models about how they navigate and work in their virtual courses. There are a few tools with a similar purpose, e.g., TADA-Ed [44] and

Moodle Data Mining Tool [45]. Both of them provide different techniques like ElWM but instructors must have certain knowledge about data mining concepts in order to use them since they are responsible for doing the phases of selection and pre-processing of attributes and the selection of algorithms and their parameter setting.

There are more specific tools, focused on a kind of problem. For example, García et al. [46] described a collaborative EDM tool based on association rule mining for the ongoing improvement of e-learning courses, allowing teachers with similar course profiles to share and score the discovered information. Kotsiantis [31] developed a decision support tool to predict students' mark from the analysis of a list of attributes defined by the user as a spreadsheet in the CSV (Comma-Separated Value) file format. The tool builds a regressor using the M5-rules algorithm [47] and ranks the influence of each attribute according to a statistical measure named RRELIEF, whose goal is to estimate the quality of attributes according to how well their values distinguish between the instances that are close to each other.

Other tools have been developed and embedded in an LMS. This is the case of the recommender integrated in the adaptive web-based educational system Aha! [48] in order to propose the most appropriate links and Web pages to students. It also includes a web mining tool to help instructors to fulfill the whole web mining process but, unlike our proposal, its use is limited to the interaction produced in the system.

15.3 E-Learning Web Miner

In this section, we describe our web-based tool, its enhanced architecture and the new services that it offers as well as its internal mode of working.

15.3.1 Description of E-Learning Web Miner

ElWM [43] is designed to help instructors to improve the teaching–learning process since it offers models and patterns that allow them to gain insight into the activity performed by their students in their virtual courses, analyze their course design and also discover their students' behavior profile and the kind of sessions they performed. In such a way, they make informed decisions driven by data. Its characteristic is that, despite using DM, end users do not require DM knowledge for its use, since the knowledge discovery process is automated and hidden. In the previous version [11], teachers only had to send a data file according to one of the templates provided by the application and request the results. As teachers considered that the task of preparing the data file was cumbersome, in the current version a new service for Moodle and Blackboard has been built, in such a way that the instructor only has to choose the question and indicate the course under study, as it is stated in Sect. 15.3.2.

Regarding DM techniques, the tool defines a template for each question and chooses the algorithm and its parameter setting according to the data file and kind of problem (association rules, clustering, classification and so on).

The techniques chosen [30, 38, 49–51] are easier to interpret and represent their results since our tool is addressed to non-expert data miners. All that means that we prefer, for instance, yacaree [52] as a rule associator rather than the well-known apriori (for instance, Borgelt implementation [53]) since the number of rules that it offers is more reduced; or kmeans or kmedoids [54] for clustering instead of a hierarchical or density-based method; and decision trees instead of neural networks, vector support machines or genetic techniques because they are more interpretable and more suitable for the size of educational datasets that, most of times, collect categorical attributes. In general, we configure these algorithms with default parameters since our experimentation shows that it is suitable for this kind of datasets.

15.3.2 General View of the E-Learning Web Miner Architecture

Our tool has been designed following a service-oriented architecture (SOA). The term SOA describes a concept to align an enterprise's IT environment with its business processes. This is achieved by providing loosely coupled atomic services that can be flexibly combined with others. A SOA can be implemented with the help of any arbitrary service-based architecture, but web services (WS) are most commonly used. We must say that our tool has been designed as a service that makes use of other services offered by our tool, though in the future, it could be orchestrated with other external services to offer a more powerful functionality. Figure 15.1 depicts this architecture.

As can be observed, we have created a web application with an easy-to-use application programming interface (API) that puts in communication the different WSs implemented. The web application, as previously mentioned, inquires the instructor about what kind of query he or she wants to make, e.g., predicting students' performance according to the global activity carried out by the students in a course or discovering who the leaders in a class are through their interaction in the forum. The API, supported by a set of configuration files where the KDD process is set up, requests its participation to each WS. Next, we describe each WS developed. The WS-ElWM Service exposes the services that can be consumed by a client application or component software that wants to include the functionality of ElWM, for example, a LMS. This service is found, when it is publicly available, or be statically bound, and possibly, choreographed into a composite service. It must be said that ElWM is an adaptation of this architecture to the e-learning context but is easily configurable for other contexts. The service is stated in the web services description language (WSDL). The WSDL file describes four data:

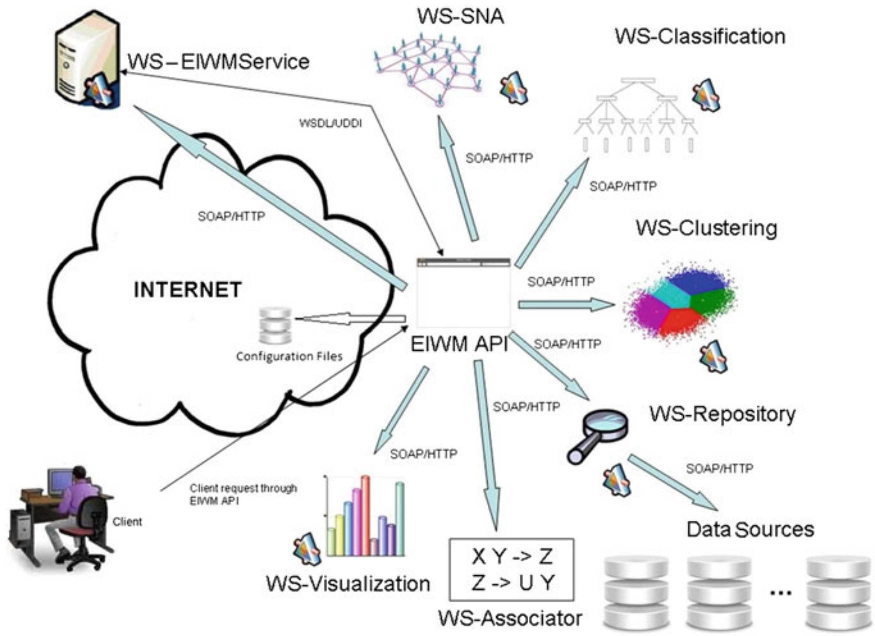


Fig. 15.1 ElWM SOA architecture

(1) interface information describing all publicly available functions (<definitions>); (2) data type information for all message requests and message responses (<types>); (3) binding information about the transport protocol to be used (<binding>), in this case Simple Object Access Protocol (SOAP); (4) address information for locating the specified service (<service>). An example can be seen in the WSDL file example presented as in the next program code.

```
<definitions xmlns:wsu"http://docs.oasis-open.org/...>
                ...
<types>
        <xsd:schema>
                <xsd:import namespace=
                "http://server/" schemaLocation="http://.../WServices?xsd=1"/>
                ...
        </xsd:schema>
</types>
                ...
<binding name="WS_RepositoryPortBinding" type="tns:WS_Repository">
        <service name="WS_Repository">
                ...
```

[WSDL file example]

Since ElWM is independent of the LMS used (Moodle, Blackboard, etc.), it is necessary to develop a service, named WS-Repository, which configures the access to the data repository of the e-learning platform in order to read the data to answer the educator question.

For each question that can be answered by ElWM, there exists an eXtended Markup Language (XML) file describing which attributes must be extracted to answer the question. As previously mentioned, ElWM answers different questions classified by the kind of data mining problems: SNA, clustering, classification, and rule association. Therefore, we have wrapped the most interpretable algorithms under each paradigm with the aim of combining more than one for each question. For example, when the tool has to do a clustering, it will use the EM algorithm to know the number of clusters and then, invoke *k* means with that number of clusters. This is the main difference of this new architecture with respect to the previous version in which each question was answered by a WS. That means that the WS had preset the DM process, whereas it is collected in an XML file now (see Sect. 15.3.4). This makes the task of improving and expanding the process easier.

Finally, in order to make it easier for the instructor to interpret the results, a WS-visualization service has been defined, whose goal is to display the results returned by the DM or SNA services graphically. For example, if the result is a classification tree, this service will make use of the Weka visualization library [41] to represent the tree graphically. Currently, we represent clusters by means of spiders and bar diagrams, and we display association rules using a MATLAB library.

15.3.3 New Services Provided

Initially, the set of models that we proposed and implemented in ElWM used only descriptive techniques, e.g., clustering and association rules, because these easily allow instructors to gain an insight into students' characteristics and depict students' learning patterns [43].

Now, we have added new services focused on classification and SNA techniques. In particular, we include statistical data about the degree of collaboration produced among learners, educators, and resources applying SNA techniques. This offers instructors the possibility of building social networks from the activity performed in forums or blogs to discover which students are more active, how the resources are used, detect isolated learners, and so on. It also provides the possibility of detecting the social communities defined in the virtual course.

Likewise, we offer the generation of models to predict the students' performance and the students' dropout because they are very useful for educational context and are very demanded by educators. ElWM allows educators to choose which type of activity performed in the LMS must be considered to build the model, e.g., getting a model based exclusively on the global activity performed in the virtual course measured by means of the time spent on each tool, the number of sessions performed, the number of messages written the replied in mail or forums,

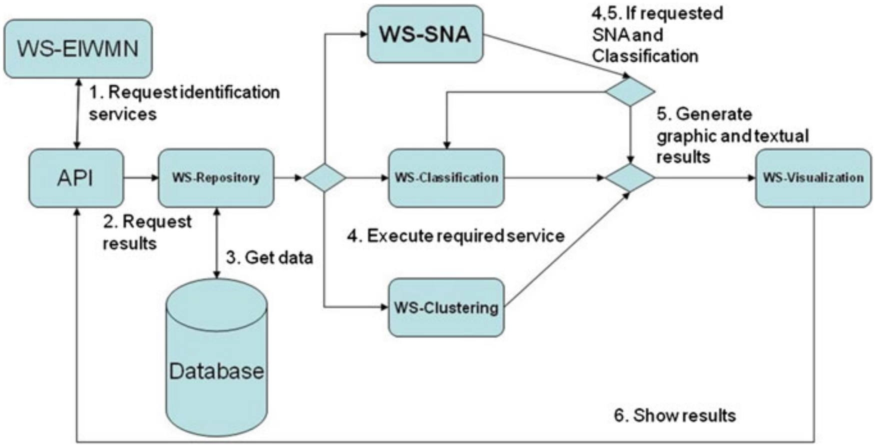

WS-ElWMN
1. Request identification services
API
WS-Repository
2. Request results
3. Get data
Database
WS-SNA
WS-Classification
WS-Clustering
4. Execute required service
4.5. If requested SNA and Classification
5. Generate graphic and textual results
WS-Visualization
6. Show results

Fig. 15.2 ElWM workflow

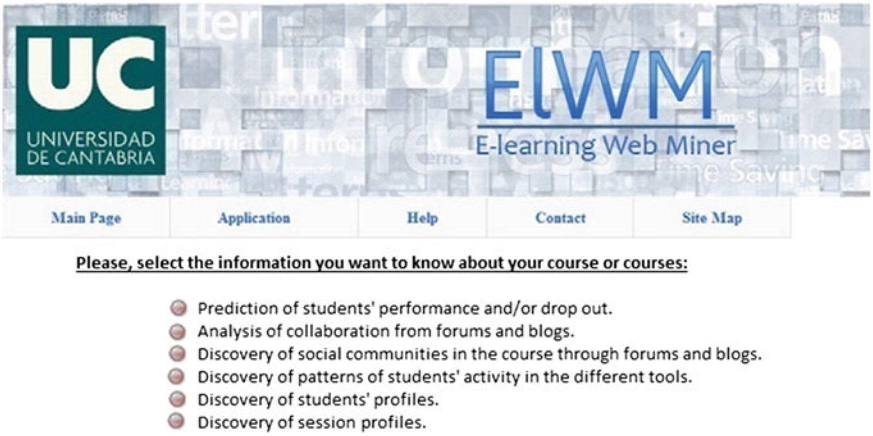
and the number of pages edited in a wiki; or getting a model that only considers the activity carried out on certain resources. Thanks to adding SNA in ElWM, these predictive models can include and take advantage of SNA features in such a way that educators can assess the effectiveness of collaboration within the group for the students to pass the course.

Currently, ElWM implements prediction models using the J48 classification algorithm, which is not only easy to interpret but also return the most accurate results, according to our experimentation [55].

In short, we have added new services focused on classification and SNA techniques. The questions that can be answered by ElWM are shown in Fig. 15.2. The new ones are "Prediction of students' performance and/or dropout" (classification task), "Analysis of collaboration from forums and blogs" (SNA), and "Discovery of social communities in the course through forums and blogs" (SNA). When an instructor selects a query, e.g., "Prediction of students' performance and/or dropout," ElWM inquires information related to this process, as can be observed in Fig. 15.3.

15.3.4 Mode of Working

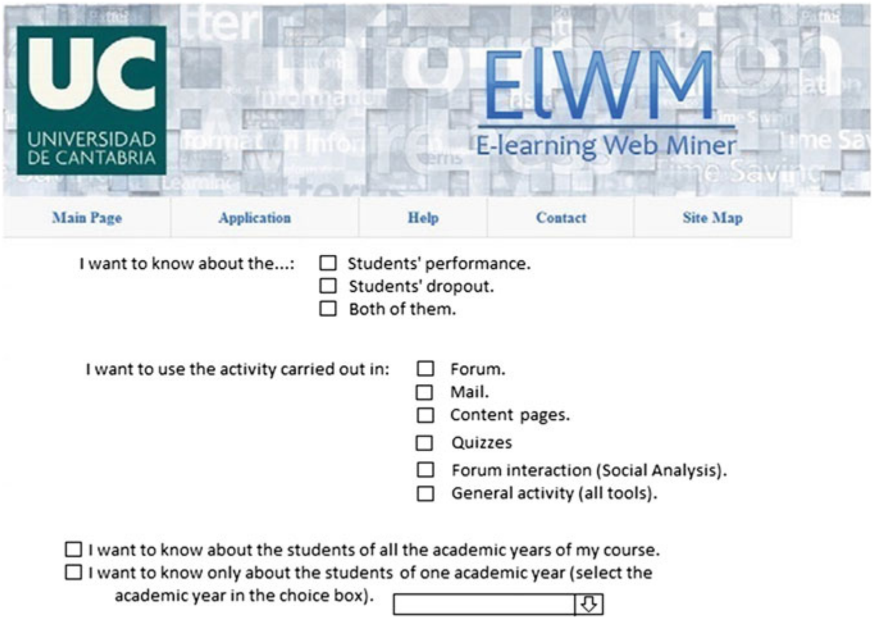
Figure 15.2 depicts how ElWM coordinates the call of the different services from the user request to the model generation and the presentation of results. The instructor through a web form (see Figs. 15.3, 15.4), will select what question he or she wants to inquire. After that, the WSs have to be identified by WSDL and then they will exchange SOAP messages between them. The first service to be called is WS-Repository, which will return the concrete data needed, obtained by querying the LMS database.



Please, select the information you want to know about your course or courses:

- ● Prediction of students' performance and/or drop out.
- ● Analysis of collaboration from forums and blogs.
- ● Discovery of social communities in the course through forums and blogs.
- ● Discovery of patterns of students' activity in the different tools.
- ● Discovery of students' profiles.
- ● Discovery of session profiles.

Fig. 15.3 ElWM questions



I want to know about the...: □ Students' performance.
□ Students' dropout.
□ Both of them.

I want to use the activity carried out in: □ Forum.
□ Mail.
□ Content pages.
□ Quizzes
□ Forum interaction (Social Analysis).
□ General activity (all tools).

□ I want to know about the students of all the academic years of my course.
□ I want to know only about the students of one academic year (select the academic year in the choice box).

Fig. 15.4 Classification options provided by ElWM

In the next step, the API will call one of the WSs among the following: WS-Classification, WS-Clustering and WS-SNA, depending on what information the instructor has requested. Note that these services are not exclusive: for example, to answer the question about the students' performance using the SNA information, ElWM has to first use the WS-SNA to generate this information and send it to the WS-Classification service in order to make the prediction task.

This last selected service will return the textual results and send them to the WS-Visualization service, which will return the same result graphically. Finally, both graphical and textual results will be shown to the instructor by the web interface (see Figs. 15.6, 15.8).

After the user has selected the options, the ElWM API has to access an XML configuration file that records, for each question and its options, which WSs are needed to get the right results. Both WS-Repository and WS-Visualization are always invoked because the former is needed to get the data from the database and the latter to visualize the results. The rest of WS are invoked depending on the questions to be answered. A piece of its content is shown in the program code XML configuration.

Such a program code depicts that, if the user selects the question to obtain the students' performance and/or dropout (*<question id="Q1"*) and he or she only wants to consider the activity carried out by the students in mail tool (*<option name="Mail"*>), then the needed WS is WS-Classification. However, if the user requests the model that requires to take the forum interaction data into account (*<option name="SNAClassification"*>), then the ElWM API must first invoke WS-SNA, and next WS-Classification.

```
<question id="Q1">
    <option name="Forum">
        <webservice n="1">WS_Classification</webservice>
    </option>
    <option name="Mail">
        <webservice n="1">WS_Classification</webservice>
    </option>
    <option name="ContentPages">


    <option name="SNAClassification">
        <webservice n="1">WS_SNA</webservice>
        <webservice n="2">WS_Classification</webservice>
    </option>
</question>
<question id="Q2">
                    ...
```

[XML configuration file]

The code SOAP messages to get the data from Moodle displays an example of the SOAP messages exchanged between the ElWM API (the requester) and the WS-Repository (requested). The first SOAP message contains the parameters (see *Body* tag) that the method *getData* of the WS-Repository needs to know in order to perform the operation of accessing the data in Moodle and returning to the API.

The parameters indicate that the data requested is for a prediction task whose class attribute is the students' performance (*"<performance>"*). Furthermore it indicates that, this task must consider the activity carried out by the students in all the tools (*"<tool>"*) from the course selected, in this case *Course1* (*"<course>"*). Finally, the parameters related to the user login are provided to connect to Moodle database.

The second message is the response that the WS-Repository sends to ElWM API after accessing the Moodle database and executing the predefined SQL sentence to get the requested data. WS-Repository returns the *courseData.xml* file that presents the format shown in the code XML data from Moodle. This process is repeated between the ElWM API and the needed WSs to get the final results.

15.4 Case Study

This section aims to show the usefulness of ElWM for the educational context. We configure different queries on several virtual courses in the same way any teacher would do and we also discuss how educators can take advantage of its use. We organize the experiments in two sections: SNA in e-learning context and prediction of students' performance and dropout.

15.4.1 Courses

First, we briefly describe the courses used for this purpose. We work with four virtual courses imparted at the UC, three of them hosted in the Moodle platform and another one in Blackboard.

The first course, hosted in Blackboard and entitled "Introduction to multimedia methods," offered in three academic years (2007–2010) with an average of 70 students enrolled from different degrees (economics, engineering, and sciences). The second course, a computer science course taught in the 2007–2008 academic year with a total of 432 enrolled students from the Computer Science degree. A course oriented to train transversal skills named "Creativity and Innovation" imparted during the first semester of 2013 with 28 learners enrolled and, finally, a congress named "Congress of Learning Styles[1]" where four discussion forums were organized.

The reasons to choose these courses were: (1) the students enrolled had different demographic profiles, (2) these learners were enrolled in degrees from different branches of knowledge, (3) the design of these courses was completely virtual and (4) we had a good relationship with the teachers to discuss the results later on.

[1] See http://congresoestilosdeaprendizaje.blogspot.com..

```
<!-- Soap Request -->
  <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
        <S:Header>
                <To xmlns="http://www.w3.org/2007/08/addressing">
                        http://localhost:8080/ElWM/WS_Repository</To>
                <Action xmlns="http://www.w3.org/2005/08/addressing">
                        http://server/WS_Repository/getData</Action>
                <ReplyTo xmlns="http://www.w3.org/2007/08/addressing">
                                ...
        </S:Header>
        <S:Body>
                <ns2:getData xmlns:ns2="http://server">
                        <prediction>performance</prediction>
                        <tool>all</tool>
                        <course>Course1</course>
                        <loginData>
                                <user>User1</user>
                                <password>pass</password>
                                <datasource>
                                        http://localhost:3535/Moodle
                                </datasource>
                        </loginData>
                </ns2:getData>
        </S:Body>
  </S:Envelope>

  <!-- Soap Response -->
  <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
        <S:Header>
                <To xmlns="http://www.w3.org/2007/08/addressing">
                        http://www.w3.org/2007/08/addressing/anonymous</To>
                <Action xmlns="http://www.w3.org/2005/08/addressing">
                        http://server/WS_Repository/getData</Action>
                                ...
        </S:Header>
        <S:Body>
                <ns2:getDataResponse xmlns:ns2="http://server">
                        <return>/.../courseData.xml</return>
                </ns2:getDataResponse>
        </S:Body>

  </S:Envelope>
```

[SOAP messages to get the data from Moodle]

```
<numberOfRows>675</numberOfRows>
   <classValueName>class</classValueName>
   <row id="1">
         <attribute name="totalTimeSpent">17239</attribute>
         <attribute name="numberOfSessions">178</attribute>
                              ...
         <attribute name="class">Pass</attribute>
   </row>
   <row id="2">
                              ...
```

[XML data from Moodle]

15.4.2 Social Network Analysis in E-Learning Courses

ElWM offers educators two questions to be answered using SNA: (1) analysis of collaboration from forums and blogs and (2) discovery of social communities in the course through forums and blogs. Our experimentation begins with the study of the course “Introduction to Multimedia Methods,” with which we made a general SNA of the students and instructors interaction in the forum. In particular, Fig. 15.5 shows the social networks for the academic year 2008–2009. For this purpose, ElWM automatically extracts the data with all the interactions of the students in the forum querying the Moodle Database at the UC. In this course, we found a single connected component and a diameter of 11, as well as low values of density (0.07) and reciprocity (7 % of the links were reciprocal).

In the previous and the subsequent courses (2007–2008 and 2009–2010), we found three and one connected components, and diameters of 19 and 16, respectively, as well as low values of density (0.05 and 0.06) and reciprocity (12 and 8 %). A possible explanation for the low values detected of both density and reciprocity is that the instructor answered the questions in the forum faster than students, preventing them from helping one another. As can be observed, the node with more links is the one corresponding to the main instructor (node 1). With this information extracted by ElWM, the instructor can know that, in this course, most interactions in the forum occur between the instructor and the students, whereas it is less frequent that those interactions that occur among the students themselves. Thus, the forum is mainly used in two different ways: (1) students make questions about the contents or the organization of the course that should be answered by the instructor and (2) the instructor makes important announcements.

These conclusions can be better understood by analyzing the node centrality values exposed in Table 15.1. As can be observed in Fig. 15.5, the instructor (node 1) has the highest values of degree and outdegree. Moreover, the difference in outdegree between the instructor and the second and the third ranked users is very high. This also happens to the betweenness and hub centrality measures. Thus, we

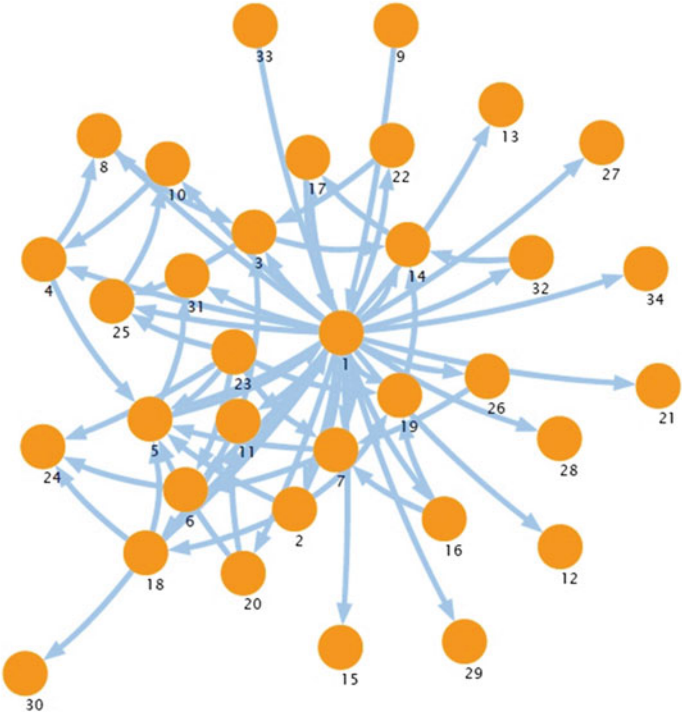


Fig. 15.5 Network of interactions between the instructor and the students of the course "Introduction to Multimedia Methods" taught in 2008–2009 at the UC

Table 15.1 Ranking of top 3 nodes for different centrality measures

	First ranked		Second ranked		Third ranked	
	Node ID	Value	Node ID	Value	Node ID	Value
Degree	1	166	3	39	5	35
Indegree	3	36	5	33	6	17
Outdegree	1	157	2	6	23	6
Betweenness	1	505	17	151	14	152
Authority	3	0.93	5	0.76	6	0.41
Hub	1	1.41	23	0.03	2	0.03

can conclude that the instructor is the user that answered the great majority of messages posted by the students in the forum.

On the other hand, the highest indegree and authority values correspond to nodes 2, 3, and 6. These students are the users that posted more messages in the forum. As a matter of fact, these three students scored the best in the course. Thus, with this analysis, we can conclude that students with a high number of interactions in the forum are likely to get good scores, a fact to be analyzed using DM techniques.

The instructor, thanks to ElWM, can obtain these results easily, with both textual and graphical visualizations. For example, ElWM does not only show the rank with the degree attribute, but it also informs the instructor about what it means. Figure 15.6 shows an example of what would be presented to the instructor in case he or she wanted to know about the interaction of students in the forum.

With this, instructors can have a better understanding of the students' social behavior and improve their participation in the course. Similar results were obtained with the other two academic years, 2007–2008 and 2009–2010.

Another example of what an instructor can obtain with ElWM, using the SNA service, is shown by means of the Congress of Learning Styles, which was celebrated in 2012. The number of people who enrolled in the course—including the 13 organizers—was 375, of which 155 actively participated in, at least, one of the 4 forums available. These forums, with which we have built the social network of interactions between organizers and participants, were used to make questions about the different topics covered by the congress, as well as answer participants' doubts and make announcements by organizers.

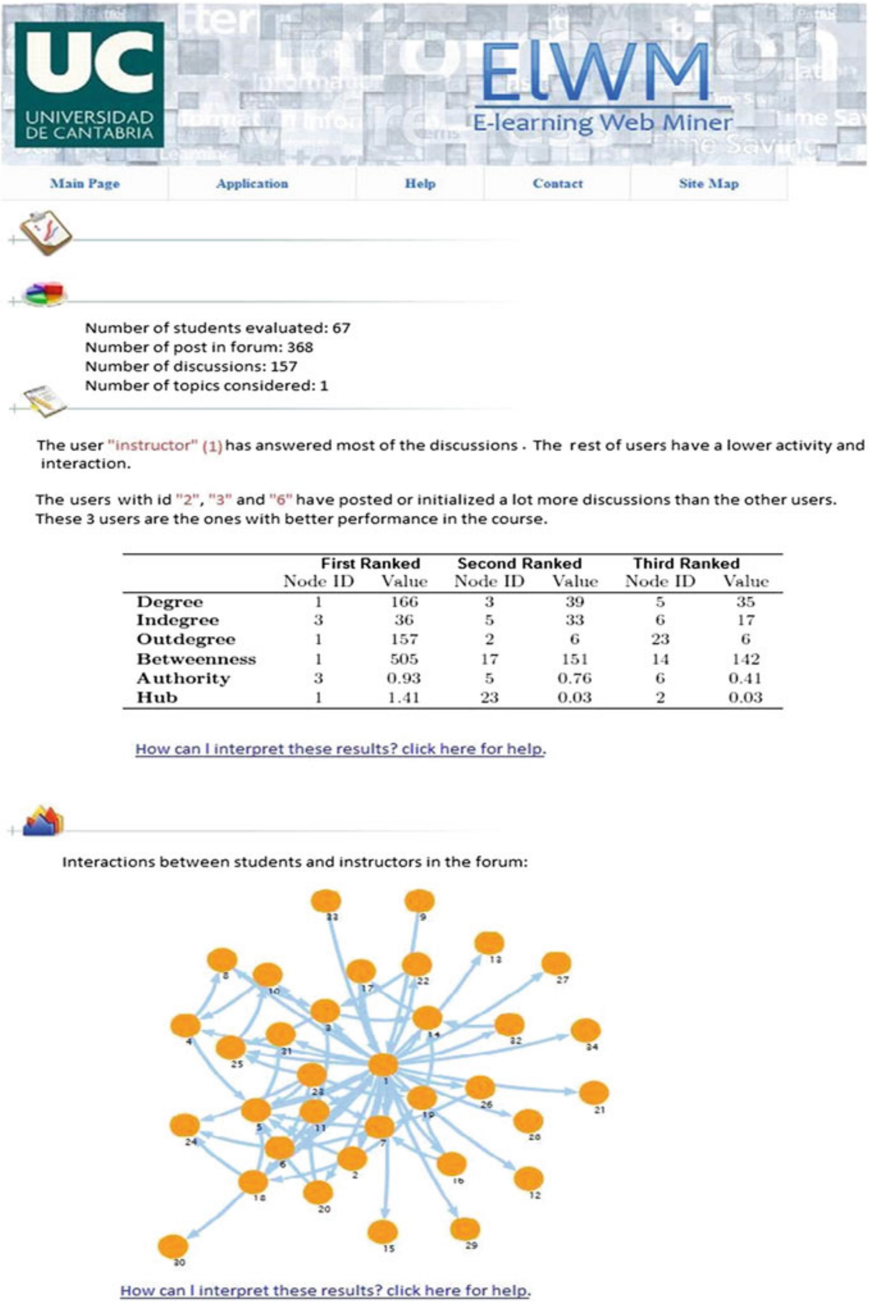
In this congress, ElWM was able to find a giant component of 155 people with a diameter of 15, as well as low values of density (0.017) and reciprocity (29 %). An explanation for such a low density is that the vast majority of people who intervened in the forum were never responded; note that just the 29 % of people who asked a question received an answer. The most important member of this forum is a student, code-named S232, who acts as a hub. As a matter of fact, in this congress, we can see how people who act as hubs, i.e., they answer the questions of people who are highly responded, are considered by SNA to be the most important actors in the network.

Although this behavior might seem logical, we have just seen how, in an educational course context, authority nodes can also turn out to be the most important ones. Also, it is noticeable that no organizer is considered to be important, i.e., most interactions in the forum occur among participants, whereas it is less frequent that those interactions occur between them and organizers.

Finally, community detection algorithms, which are able to identify clusters of vertices with a high internal density of edges, were also considered to check whether the forum structure of the congress is independent and recognizable enough to be recovered by such a type of algorithms.

For this purpose, FRINGE [12], an overlapping community detection algorithm with a recent approach based on the dynamics of social networks, was wrapped in ElWM and used. It was able to recover four overlapping community members, i.e., S232, S303, S296, and S220, as leading members of every community, respectively. As can be observed in Fig. 15.7, these leading members are the same as those obtained in the four forums of the congress.

This result demonstrates the power of community detection algorithms and justifies their use under this kind of scenarios. The educational context would allow instructors to discover the students' degree of interaction, which is paramount to



Number of students evaluated: 67
Number of post in forum: 368
Number of discussions: 157
Number of topics considered: 1

The user "instructor" (1) has answered most of the discussions . The rest of users have a lower activity and interaction.

The users with id "2", "3" and "6" have posted or initialized a lot more discussions than the other users. These 3 users are the ones with better performance in the course.

	First Ranked		Second Ranked		Third Ranked	
	Node ID	Value	Node ID	Value	Node ID	Value
Degree	1	166	3	39	5	35
Indegree	3	36	5	33	6	17
Outdegree	1	157	2	6	23	6
Betweenness	1	505	17	151	14	142
Authority	3	0.93	5	0.76	6	0.41
Hub	1	1.41	23	0.03	2	0.03

How can I interpret these results? click here for help.

Interactions between students and instructors in the forum:

How can I interpret these results? click here for help.

Fig. 15.6 ElWM results applying SNA to the course "Introduction to Multimedia Methods" taught in 2008–2009 at the UC

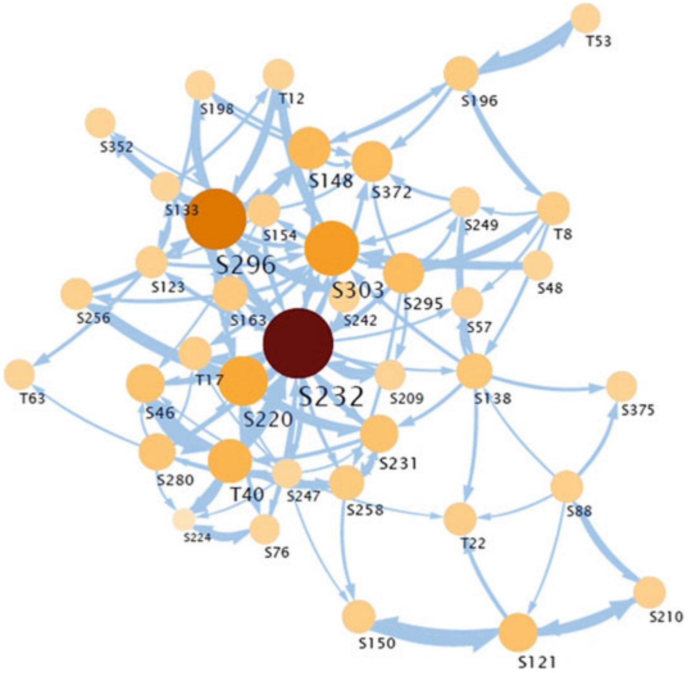


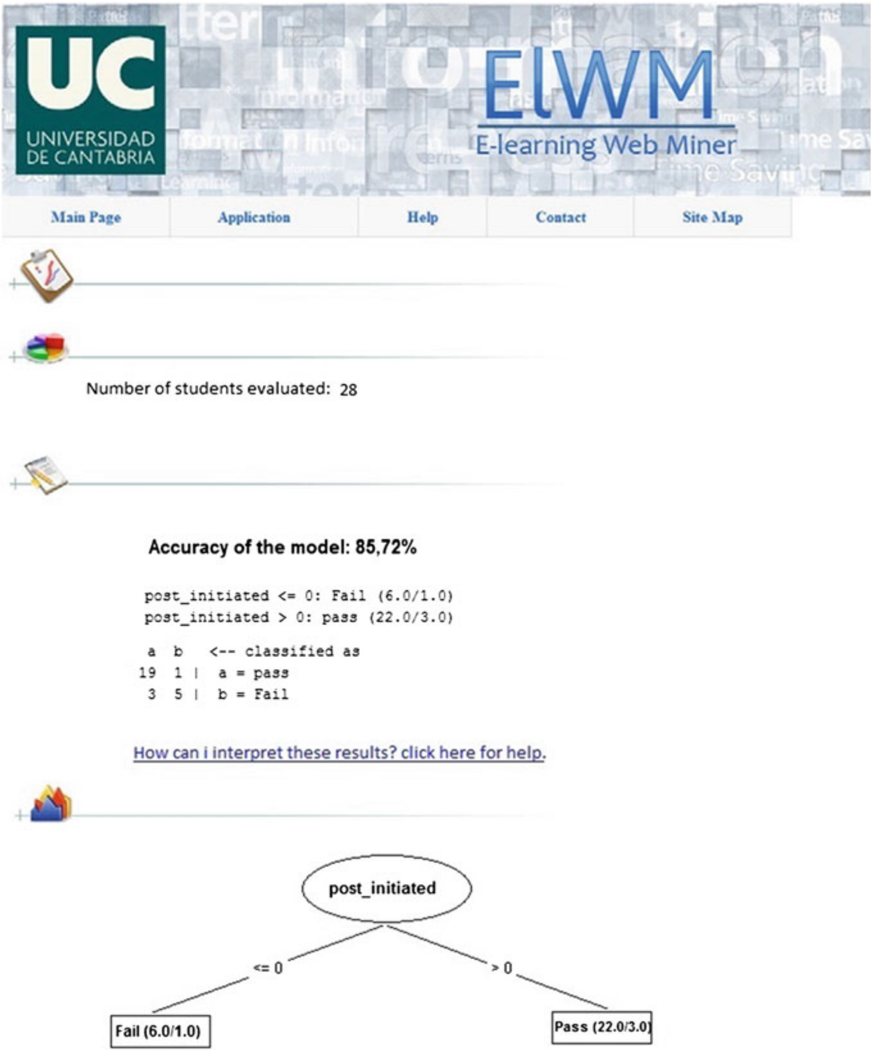
Fig. 15.7 Network of interactions among the participants in "Congress of Learning Styles"

create team-works with different social profiles. Also, it could be used to score the contribution of each student to a certain forum's discussion. Thus, it would take an objective variable of participation into consideration.

15.4.3 Prediction of Students' Performance and Dropouts

As shown in Fig. 15.3, ElWM provides educators with several configuration options to generate models and patterns. First of all, instructors must choose the kind of model: performance, dropout or both, and select the tools that must be analyzed according to the design of the course and the goal of their analysis. Furthermore, they can determine if the analysis is carried out for one specific academic year or for all of them.

Our first example shows the pattern obtained by ElWM for the "Creativity and Innovation" course when forum resource was chosen. This implies the use of the following attributes: number of posts read, number of posts initiated, and number of messages posted by each student in order to discover whether learners will pass. Figure 15.8 displays the result shown to the end user.



Number of students evaluated: 28

Accuracy of the model: 85,72%

```
post_initiated <= 0: Fail (6.0/1.0)
post_initiated > 0: pass (22.0/3.0)

 a  b   <-- classified as
19  1 |  a = pass
 3  5 |  b = Fail
```

How can i interpret these results? click here for help.

Fig. 15.8 ElWM classification results for "Creativity and Innovation" course

This is a tree, obtained with J48 algorithm from Weka set up with default parameters, which says that students, who initialize 1 or more posts, would pass the course with 86 % of accuracy and students who do not initialize any posts would fail with an accuracy of 95 %. The global accuracy of the model is 85.72 %. Hence, the instructor could encourage their learners to participate in forums because the information written was useful to pass the course.

A more complex and rich output would be obtained if the number of students and resources to be analyzed grew. It is partially shown in program code for the view of the classification result for "Computer Science" course, where the teacher responsible for the Computer Science course selected the options to build a students' performance model taking the global activity into account, i.e., the total time spent on the course and total number of sessions performed, the forum activity measured from the number of posts read, number of posts initiated and number of messages posted by each student, the mail activity with the number of mails sent and received and the number of quizzes performed. In this case, the model achieves an accuracy of 70.12 %.

```
n_quiz_a <= 5
    n_assignment <= 1
        n_posts <= 0: fail
        n_posts > 0
            total_time_quiz <= 8164
                total_time_assigment <= 217: fail
                total_time_assigment > 217: pass
            total_time_quiz > 8164: pass
    n_assignment > 1
        n_read <= 1
            n_assignment <= 6
                n_assignment <= 2
                    total_time_forum <= 555: pass
                    total_time_forum > 555
                ...
```

[Partial view of the classification result for "Computer Science" course]

It also indicates that the number of quizzes completed is the most important students' activity to be considered in order to classify their performance, but other attributes, such as the total time spent on the forum and the number of submitted assignments is also relevant. Finally, we present an example in which data from SNA is used. We show the results obtained from the Multimedia course used in Sect. 15.4.2. In this case, the goal was to predict performance and dropout using the following global activity attributes (in this case, since this course was imparted in Blackboard, the number of global attributes that ElWM can get is higher): total time spent, total number of sessions, average time spent per week and average number of sessions per week and the activity in forum and mail.

The model generated is shown in the code for Partial view of the classification model with SNA attributes. By reading this code, the teacher discovers that students with an average time per week lower than 63 min are likely to dropout, and also finds out that the interaction among students in the forum is important for the classification task: the students with an average time per week equal or higher than 63 min are considered to fail or pass the course depending, for example, on whether they are an authority in the forum or if they have written a high quantity of posts. By using ElWM and having these last results, the instructor can know not only about the interactions of the students, but also how these interactions affect the prediction of students' performance and dropout.

15.5 Conclusions

In this chapter we describe the architecture and functionality of an educational tool that assists educators to monitor, analyze, and better understand the behavior of their students in the development of collaborative activities using Web 2.0 resources.

```
average_time_per_week <= 63: dropout
average_time_per_week > 63
    number_of_messages_written_in_the_forum <= 0
        average_number_of_sessions_per_week <= 3
            number_of_messages_read_in_the_forum <= 52: pass
            number_of_messages_read_in_the_forum > 52: dropout
        average_number_of_sessions_per_week > 3
            total_time <= 1962: pass
            total_time > 1962
                total_time <= 2000: dropout
                total_time > 2000: pass
    number_of_messages_written_in_the_forum > 0
        number_of_messages_written_in_the_forum <= 8
            in_degree_centrality <= 0.013
                number_of_messages_read_in_the_forum <= 87: pass
                number_of_messages_read_in_the_forum > 87
                    number_of_messages_read_in_the_mail <= 24
                    number_of_messages_written_in_the_forum <= 4
            in_degree_centrality_unbalanced <= 1
                authority_centrality <= 0.029
                    number_of_messages_read_in_the_mail <= 7: fail
```

[Partial view of the classification model with SNA attributes]

For instances forums, blogs, or wikis, as well as, achieving course objectives. This tool, called E-learning Web Miner, relies on the application of DM and SNA techniques on data from logs of the services used to develop the teaching–learning process (e-learning platforms, social networks, wiki spaces, etc.).

Its architecture, based on WSs, makes it an easily extensible and embeddable one in any tool, as for example in an LMS. Its main feature is that its use is oriented to non-expert educators in analytics since it hides the knowledge discovery processes from the user. Its mode of working is simple; instructors have to connect to ElWM and choose the question to solve by pointing out the virtual courses under study.

ElWM carries out the mining process without user's interaction and displays the results in textual and graphical mode. This new version of ElWM includes three new questions: "Prediction of students performance and/or dropout," "Analysis of collaboration from forums and blogs," and "Discovery of social communities in the course through forums and blogs" using classification techniques and social analysis techniques for these tasks.

Currently, our research is focused on meta-learning [30] to build a recommender that automatists the choice of the most suitable classification algorithm for each dataset at hand. Furthermore, we are adapting our FRINGE algorithm to directed and weighted networks with the aim of finding out the leaders (our most active learners) in a network according to their weight (e.g., messages answering doubts in a forum can have less value than contributions in a wiki from an instructional point of view).

In addition to the direction of the edges, i.e., the relationship between learners (both support themselves or the help is always in one sense). At the same time, we are working in the phase of tool testing and we hope to deploy it in the coming months. Next, we will study new and interesting questions for educators and look for the best answers to these questions. This will lead us to add new algorithms and visualization tools in our web services.

Acknowledgements The authors would like to thank the anonymous referees for their constructive comments, which led to a significant improvement of this paper. The authors are also deeply grateful to CEFONT, the department of the UC that is responsible for LCMS maintenance, for their help and collaboration. Likewise, the authors gratefully acknowledge the valuable collaboration of the instructors involved in the courses analyzed. This work is partially supported by the Ministry of Education, Culture, and Sport of the Government of Spain, under grant Beca de colaboración (2012–2013), and by the UC, under a Ph.D. studentship (2011–2015).

References

1. Capra, R., Arguello, J., Chen, A., Hawthorne, K., Marchionini, G., Shaw, L.: The results space collaborative search environment. In: 12th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 435–436. ACM, New York (2012)
2. Lin, C.C., Tsai, C.C.: Participatory learning through behavioral and cognitive engagements in an online collective information searching activity. Int. J. Comput. Support. Collaborative Learn. **7**(4), 543–566 (2012)
3. McNely, B.J., Gestwicki, P., Hill, J.H., Parli-Horne, P., Johnson, E.: Learning analytics for collaborative writing: a prototype and case study. In: Dawson, S., Haythornthwaite, C., Shum, S.B., Gasevic, D., Ferguson, R. (eds.) Second International Conference on Learning Analytics and Knowledge, pp. 222–225. ACM, New York (2012)
4. Joubert, M., Wishart, J.: Participatory practices: lessons learnt from two initiatives using online digital technologies to build knowledge. Comput. Educ. **59**(1), 110–119 (2012)
5. Rice, W.: Moodle E-learning Course Development. A Complete Guide to Successful Learning Using Moodle. Packet Publishing, Birmingham (2006)

6. Southworth, H., Cakici, K., Vovides, Y., Zvacek, S.: Blackboard for Dummies. Wiley, New York (2006)
7. Korcuska, M., Berg, A.M.: Sakai Courseware Management: The Official Guide. Packt Publishing, Birmingham (2009)
8. Johnson, L., Adams-Becker S., Cummins, M., Estrada, V., Freeman, A., Ludgate, H.: NMC Horizon Report: Higher Education Edition. Report. The New Media Consortium (2013)
9. Romero, C., Ventura, S.: Data mining in education. Wiley Interdiscip. Rev.: Data Min. Knowl. Disc. **3**(1), 12–27 (2013)
10. Macfadyen, L.P., Dawson, S.: Mining LMS data to develop an early warning system for educators: a proof of concept. Comput. Educ. **54**(2), 588–599 (2010)
11. Zorrilla, M., García-Saiz, D.: A service oriented architecture to provide data mining services for non-expert data miners. Decis. Support Syst. **55**(1), 399–411 (2013)
12. Palazuelos, C., Zorrilla, M.E.: FRINGE: a new approach to the detection of overlapping communities in graphs. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011. LNCS, vol. 6784, pp. 638–653. Springer, Heidelberg (2011)
13. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: an issue brief. U.S. Department of Education, Office of Educational Technology (2012)
14. Scott, J.: Social Network Analysis: A Handbook. SAGE Publications, London (2000)
15. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge (1994)
16. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature **393**(6684), 440–442 (1998)
17. Klovdahl, A., Potterat, J., Woodhouse, D., Muth, J., Muth, S., Darrow, W.: Social networks and infectious disease: the colorado springs study. Soc. Sci. Med. **38**(1), 79–88 (1994)
18. Krebs, V.: Mapping networks of terrorist cells. Connections **24**(3), 43–52 (2002)
19. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
20. Kleinberg, J.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
21. Brewe, E., Kramer, L.H., Sawtelle, V.: Investigating student communities with network analysis of interactions in a physics learning center. Phys. Rev. Spec. Top. Phys. Educ. Res. **8**(1), 1–9 (2012)
22. Crespo, P.M.T., Antunes, C.: Social networks analysis for quantifying students performance in teamwork. In: Yacef, K., Zaïane, O., Hershkovitz, A., Yudelson, M., Stamper, J. (eds.) 5th International Conference on Educational Data Mining, pp. 234–235. International Educational Data Mining Society, Chania (2012)
23. Cuéllar, M.P., Delgado, M., Pegalajar, M.C.: Improving learning management through semantic web and social networks in e-learning environments. Expert Syst. Appl. **38**(4), 4181–4189 (2011)
24. Rabbany, R., Takaffoli, M., Zaïane, O.R.: Social network analysis and mining to support the assessment of on-line student participation. SIGKDD Explor. **13**(2), 20–29 (2011)
25. Dawson, S., Tan, J.P.L., McWilliam, E.: Measuring creative potential: using social network analysis to monitor a learners' creative capacity. Australas. J. Educ. Technol. **27**(6), 924–942 (2011)
26. Dawson, S.: Seeing the learning community: an exploration of the development of a resource for monitoring online student networking. Br. J. Educ. Technol. **41**(5), 736–752 (2010)
27. Obsivac, T., Popelinsky, L., Bayer, J., Geryk, J., Bydzovska, H.: Predicting drop-out from social behaviour of students. In: Yacef, K., Zaïane, O., Hershkovitz, A., Yudelson, M., Stamper, J. (eds.) 5th International Conference on Educational Data Mining, pp. 103–109. International Educational Data Mining Society, Chania (2012)
28. Palazuelos, C., García-Saiz, D., Zorrilla, M.: Social network analysis and data mining: an application to the e-learning context. In: International Conference on Computational Collective Intelligence Technologies and Applications (2013, in press)

29. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. J. Artif. Intell. **17**(1), 37–54 (1996)
30. García-Saiz, D., Zorrilla, M.E.: Towards the development of a classification service for predicting students' performance. In: D'Mello, S.K., Calvo, R.A., Olney, A. (eds.) 6th International Conference on Educational Data Mining, pp. 318–319. International Educational Data Mining Society, Memphis (2013)
31. Kotsiantis, S.B.: Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. J. Artif. Intell. **37**(4), 331–344 (2012)
32. Romero, C., Ventura, S., Espejo, P.G., Hervás, C.: Data mining algorithms to classify students. In: Baker, R.S.J.D., Barnes, T., Beck, J.E. (eds.) 1st International Conference on Educational Data Mining, pp. 8–17. International Educational Data Mining Society, Montreal (2008)
33. Zafra, A., Romero, C., Ventura, S.: Predicting academic achievement using multiple instance genetic programming. In: Ninth International Conference on Intelligent Systems Design and Applications, pp. 1120–1125, IEEE, Washington (2009)
34. Dekker, G., Pechenizkiy, M., Vleeshouwers, J.: Predicting students drop out: a case study. In: Barnes, T., Desmarais, M., Romero, C., Ventura, S. (eds.) 2nd International Conference on Educational Data Mining, pp. 41–50. International Educational Data Mining Society, Cordoba (2009)
35. Kotsiantis, S.B., Pierrakeas, C., Pintelas, P.E.: Preventing student dropout in distance learning using machine learning techniques. In: Palade, V., Howlett, R.J., Jain, L.C. (eds.) KES. LNCS, vol. 2773, pp. 267–274. Springer, Heidelberg (2003)
36. John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345. Morgan Kaufmann, San Francisco (1995)
37. Hämäläinen, W., Vinni, M.: Comparison of machine learning methods for intelligent tutoring systems. In: Ikeda, M., Ashley, K., Chan, T.W. (eds.) Intelligent Tutoring Systems. LNCS, vol. 4053, pp. 525–534. Springer, Heidelberg (2006)
38. Zafra, A., Ventura, S.: G3P-MI: a genetic programming algorithm for multiple instance learning. Inf. Sci. **180**(23), 4496–4513 (2010)
39. Friedman, N., Geiger, D., Goldszmidt, M., Provan, G., Langley, P., Smyth, P.: Bayesian network classifiers. Mach. Learn., 131–163. Kluwer Academic Publishers, Boston (1997)
40. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco (1993)
41. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
42. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**(6), 601–618 (2010)
43. García-Saiz, D., Zorrilla, M.E.: E-learning web miner: a data mining application to help instructors involved in virtual courses. In: Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., Stamper, J. (eds.) 4th International Conference on Educational Data Mining, pp. 323–324. International Educational Data Mining Society, Eindhoven (2011)
44. Benchaffai, M., Debord, G., Merceron, A., Yacef, K.: TADA-ED, a tool to visualize and mine students' online work. In: McKay, E., Collis, B. (eds.) International Conference on Computers in Education, pp. 1891–1897. RMIT, Melbourne (2004)
45. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Comput. Educ. **51**(1), 368–384 (2008)
46. García, E., Romero, C., Ventura, S., de Castro, C.: A collaborative educational association rule mining tool. Internet High. Educ. **14**(2), 77–88 (2011)
47. Holmes G., Hall, M., Frank, E.: Generating rule sets from model trees. In: 12th A.J.C. on Artificial Intelligence. LNCS, vol. 1747, pp. 1–12. Springer, Heidelberg (1999)
48. Romero, C., Ventura, S., Zafra, A., de Bra, P.: Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. Comput. Educ. **53**(3), 828–840 (2009)

49. Balcázar, J.L., Tîrnauca, C., Zorrilla, M.E.: Filtering association rules with negations on the basis of their confidence boost. In: International Conference on Knowledge Discovery and Information Retrieval, pp. 263–268, INSTICC, Valencia (2010)
50. Zorrilla, M.E., García, D.: A data mining service to assist instructors involved in virtual education. In: Zorrilla, M., Mazón, J., Ferrández, Ó., Garrigós, I., Daniel, F., Trujillo, J. (eds.) Business Intelligence Applications and the Web: Models, Systems and Technologies, pp. 222–243. Business Science Reference, Hershey (2012)
51. Zorrilla, M.E., García-Saiz, D., Balcázar, J.L.: towards parameter-free data mining: mining educational data with Yacaree. In: Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., Stamper, J. (eds.) 4th International Conference on Educational Data Mining, pp. 363–364. International Educational Data Mining Society, Eindhoven (2011)
52. Balcázar, J.L.: Parameter-free association rule mining with Yacaree. In: Khenchaf, A., Poncelet, P. (eds.) Extraction et Gestion des Connaissances, pp. 251–254. Hermann, Brest (2011)
53. Borgelt, C.: Efficient implementations of Apriori and Eclat. In: Goethals, B., Zaki, M.J. (eds.) ICDM Workshop of Frequent ItemSet Mining Implementations. CEUR-WS, Melbourne (2003)
54. Park, H.S., Jun, C.H.: A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. **36**(2), 3336–3341 (2009)
55. García-Saiz, D., Zorrilla, M.: Comparing classification methods for predicting distance students' performance. In: Diethe, T., Balcázar, J.L., Shawe-Taylor, J., Tîrnauca, C. (eds.) Journal of Machine Learning Research, Workshop and Conference Proceedings. 2nd Workshop on Applications of Pattern Analysis, vol. 17, pp. 26–32 (2011)