Mining Student Data To Characterize Similar Behavior Groups In Unstructured Collaboration Spaces

Luis Talavera [1] and **Elena Gaudioso** [2]

Abstract. The adoption of Learning Management Systems to create virtual learning communities is a unstructured form of allowing collaboration that is rapidly growing. Compared to other systems that structure interactions, these environments provide data of the interaction performed at a very low level. For assessment purposes, this fact poses some difficulties to derive higher lever indicators of collaboration. In this paper we propose to shape the analysis problem as a data mining task. We suggest that the typical data mining cycle bears many resemblances with proposed models for collaboration management. We present some preliminary experiments using clustering to discover patterns reflecting user behaviors. Results are very encouraging and suggest several research directions.

1 INTRODUCTION

Computers and the Internet are widely used in educational contexts. Particularly, the adoption of Learning Management Systems (LMS) is rapidly growing as valuable tools for developing on-line courses. These platforms offer a great variety of channels and workspaces to facilitate information sharing and communication between participants in a course and therefore enable collaborative learning. In this context, we can identify two contrasting approaches regarding the degree of structure of the collaboration space. The first strategy typically found in the Computer Supported Collaborative Learning (CSCL) literature, relies in supporting collaboration by offering a context aware interface intended to structure the interaction. Common examples are computer mediated conversations through the use of dialog tags or sentence openers. On the other side of the spectrum, we have unstructured collaboration spaces that group course participants and offer an open interface for communication and sharing of knowledge and experiences.

The latter approach is the basis of the so called *virtual communities*, groups of people with common interests or goals that use the Internet resources to improve their communication and coordination. In education, virtual communities share a common goal of learning and are usually monitored by a tutor. A simple but commonly used pedagogical model consists of making a tutor and a set of students members of a web-based workgroup with course materials available online (and possibly also offline) and one or several shared services. Some views may consider this approach as a too weak form of a collaborative learning space when compared to more structured interfaces that support collaboration. A discussion on the pedagogical and practical convenience of either approach is outside the scope of this paper. However, we remark two aspects to support our interest in this approach. First, although providing less elaborated support, educational virtual environments enable teachers to set up structured collaborative activities. Secondly, these sort of settings are rapidly becoming popular because of the availability of powerful open source LMS allowing a relatively easy set up of the collaborative space and the flexibility they provide.

2 THE CASE FOR DATA MINING

While the amount of structure imposed to an interface might have more or less pedagogical impact, it clearly influences the task of a course instructor. First, open environments place all the burden of guiding collaboration on the instructors who have all the responsibility of enabling and controlling how the collaborative activities take place. Secondly, using structured interfaces, the information gathered from the students interaction has a higher quality and a more straightforward interpretation for assessment purposes. Powerful LMS are database backed, i.e., they rely on a full-fledged RDBMS to store not only information contents but also to track all the the interaction performed in the workspaces. Tracking logs information at a low and generic level ('user u sent message m to forum f'). To make sense of this data, it is necessary to formulate some queries to obtain aggregated results ('number of messages sent by user u to forum f'). This sort of reporting sometimes referred to as *quantitative analysis* [1]. A *qualitative analysis* step can produce a set of indicators identifying criteria to evaluate the learning process by attributing semantics to events in the workspace. Semi-structured interfaces facilitate this step by gathering high-level data such as the type of contribution to a conversation (proposal, agreement), but being generic systems, LMS do not originally provide this information. From this discussion, two research directions arise, supporting teachers in analyzing and assessing student interactions and get the best possible results out of a large amount of low-level data with weak semantics.

A model for characterizing systems that analyze interactions to support collaborative learning has been proposed in [5], where three types of supporting systems are described, namely mirroring, metacognitive and advising tools. In [6] an extended version of this model is proposed by clearly differentiating between two goals, scaffolding and evaluation. The objective of evaluation systems is to perform an analysis of collaboration to report to a student or an instructor. Usually, evaluation is performed at the end of an activity to gain an insight of how it has been developed, while scaffolding aims to real-time diagnosis and correction of problems. But perhaps, the main difference lies in that evaluation tools do not necessarily require a normative model of interaction to be useful, rather they can serve as an exploratory tool. Under the virtual community model, our first

[1] Dept. Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Diagonal 647, p8, 08028 Barcelona, Spain, talavera@lsi.upc.es

[2] Dept. Inteligencia Artificial, E.T.S.I. Informática, UNED, Juan del Rosal 16, 28040 Madrid, Spain, elena@dia.uned.es

research direction can be interpreted along these lines.

The second challenge is related to exploit the data gathered by the LMS regardless the limited semantic information available. We propose to formulate this problem as a data mining task, i.e., as a process of exploring and analyzing data to identify useful patterns. Obtained patterns are useful pieces of knowledge that can help to perform a qualitative analysis, even when the only source of information is low-level interaction data. Data mining can be viewed as a cycle that consists of several steps [3]:

- **Identify** a problem where analyzing data can provide value.
- **Collect** the data.
- **Preprocess** the data to obtain a clean, mineable table.
- **Build a model** that summarizes patterns of interest in a particular representational form.
- **Interpret/Evaluate** the model.
- **Deploy** the results incorporating the model into another system for further action.

Perhaps not surprisingly, this classical framework for data mining bears a strong resemblance with the collaboration management cycle described in [5] and [6]. The data collection phase is the same. Constructing a model of interaction involves selecting high-level variables, termed indicators, to represent the state of interaction. This phase corresponds to the data preprocessing and model building stages. The result of applying data mining is a model that summarizes patterns of interaction. The analysis phase where a diagnosis or understanding is obtained is analogous to the model interpretation/evaluation step. However, the latter may suggest modifications in the previous steps to produce alternative models. Finally, the correction phase that take into account the interaction model to perform remedial actions has a connection with the deployment stage.

Selecting high-level variables is not absolutely required in data mining because the high level information is the model itself. Nevertheless, it can be done in order to obtain more interpretable models. This additional step is a remarkable difference, since, after data preparation and transformation is finished, evaluation CSCL systems already provide a result in terms of reports. The responsibility for extracting useful patterns is left to the teacher. At this stage of the process, we could generate also a set of reports, but we go a step beyond by building a model to automatically discover patterns.

The aim of this paper is to explore the application of data mining to the data provided by a database backed LMS and build analytical models summarizing interaction patterns. The discovered model is then presented to the teacher to provide insight into the student behavior, therefore supporting interaction assessment. Specifically, we employ clustering techniques that automatically discover useful groups from data to obtain profiles of student behaviors. Additionally, by describing the whole data mining process, we aim to show that the definition of this task correlates well with the collaboration management cycle, thus suggesting a promising line of research.

3 DATA COLLECTING AND PREPROCESSING

The course from which we collected the data was aimed to teach the use of the Internet in education. We set up a workgroup for the students and instructors providing forums, news services, chats, file storage areas and personal web-pages areas. At the beginning of the course, we presented the students with a survey in order to evaluate how familiar they were with educational software, the Internet and several computer applications. Additionally, they could fill another survey indicating their interests (distance education, psychology, pedagogy,web design, etc.). In most cases the students were in turn lecturers interested in the use of the Internet in education, and on the whole, they had little experience in the use of computers and Internet services.

The course instructors closely monitored the students solving, on demand, the difficulties that arose. They proposed several collaborative activities to the students, mainly forum discussions and promoted using the platform services to help their peers. These activities were a part of a larger course project used to assess students, but collaboration was not mandatory. Although there was also the possibility of contacting the instructors by other means, the main channels of communication were the course forums and electronic mail. The instructors also visited the chat room established for the course to detect if the students had difficulties in the course. Given this scenario, our goal is to detect patterns of interaction and relate these patterns to student performance, reflected in the final course grades.

The platform for the course was build upon the Ars Digita Community System (ACS), an open source collection of applications designed to support web communities. The software provides a collection of modules for user/groups management, content management, news, FAQs, calendar, forums, etc. integrated via one single collection of database tables that refer to one another in a relational database. This database does not only maintains the personal information of the users and the contents they send, but also serves to track and structure all the events in the collaborative space. The approach of collecting data at the application server layer is typically followed by a number of LMS systems.

As usual when analyzing transactional data, the set of tables was not yet in a form suitable for mining purposes. In order to optimize transactional operations of adding, removing and updating information, relational data models are highly normalized (breaked-down) into several tables (more than 50 in our case). However, data mining algorithms require a single table containing all the information relevant for the analysis and organized at a particular level of detail (e.g. student). This table is the result of a number of preparations through selecting, grouping, pivoting or joining operations plus a data cleaning and transformation step.

In our case, the data was preprocessed using SQL queries on the database and with the help of additional data preparation software. Since our aim was to obtain data to derive group behavior profiles, we oriented our preparation process to aggregate data for each individual student from the information distributed across several tables in the database. After the resulting table was obtained, we cleaned columns with a single value. They are usually either information that no student has filled or less popular services that nobody has used for which a default value has been inserted.

We note that the most of the CSCL literature on supporting collaboration does not make reference to the underlying data model or the complexity of the queries required to obtain aggregated student information. This might be because structured interfaces rely on a single or few collaborative services, and store the data in a relatively denormalized data model already prepared to easily serve most of the projected queries. Working with a LMS using a generic and normalized data model for several communication services, we do not have this luxury and the preprocessing step is mandatory.

Interaction data can be enriched with any other possibly relevant information collected. We considered two additional types of data. In the first group we included demographic data and background knowledge obtained from surveys. The latter, included data about user interests from additional surveys. Table 1 shows interaction data.

Number of sessions started
Number of entrances to the course chat
Number of messages sent to the course chat
Number of messages sent to the course forum
Number of files in the file storage area of the course
Activated alerts in the forum of the course
Created forum
Published a web presentation
Added bookmarks
Sent a email to the whole course
Registered in other courses
Number of messages sent to other forums
Number of entrances to other chats
Number of messages sent to other chats
Number of files in other file storage areas
Activated alerts in other forums
Number of static course pages visited
Number of threads started in the course forum
Number of threads replied by other students in the course forum
Number of threads finished in the course forum

Table 1. Features related to the user interaction with the system

Including some features was more or less straightforward, since they basically reflected either the number of times an activity was performed or whether a given service was used. As we have mentioned before, this information is well suited for quantitative analysis and our aim is to deliver more elaborated knowledge from the patterns derived from a data mining model. Nevertheless, despite the unstructured setting of the course, we have attempted to derive some more elaborated features from our data taking into account the pedagogical model used in the courses. Since the main collaborative activities proposed were developed in the forums, we considered this service a good source of information to derive higher-level features. Particularly, we extracted the three last features in Table 1. Although they are still quantitative accounts of behavior, they can be attached more subjective meanings. We interpret the number of threads started as an indication of the degree of involvement to produce a contribution (initiative). Additionally, the number of student messages replied suggested a measure of how he/she is promoting discussion. We believe that this interpretation makes sense in the context of the course studied, given the voluntary nature of the collaborative activities.

Finally, most of the information collected is numerical and using it in this form would result in numerical descriptions of the group characteristics. As opposed to other type of numerical information related to concepts such as grades that is easy to interpret, data reflecting averages of the number of times someone has accessed a service results difficult to understand. A solution is to discretize those numerical features into ranges (such as low, medium, high) that provide a much more comprehensible view of the data for an average person. This process was performed manually by an instructor of the course by observing the histograms of each feature. A side-effect of this process was the removal of some features for which the instructor inferred that the behavior was more or less uniform across all the students, so that they provided no useful information. (this removal is already reflected in Table 1). A similar step is also performed in DEGREE [1] using fuzzy logic techniques, although in this case, it is used to generate final reporting indicators.

4 BUILDING A CLUSTERING MODEL

There exists a large number of clustering algorithms in the literature and the choice depends on the particular application. For our purposes, we require an algorithm capable of dealing with discrete data. Model-based clustering is an approach that has gained wide popularity in the literature for both continuous and discrete data [8]. This approach assumes that the data has been drawn from one of k sources corresponding to the clusters. We define the set of model parameters $\Theta = \{\lambda_k, \theta_i^{(k)}\}$ indicating the probability of each cluster and the distribution of each feature i in cluster k.

The parameters θ_i depend on the distribution assumed for the features. We assume a very simple but widely used model closely related to the Naive Bayes model for classification in which all the features are treated as conditionally independent given the cluster value. We model each feature with a multinomial distribution where the parameters for a given cluster are the conditional probability of each feature value given the cluster.

In this probabilistic framework, the clustering task can be viewed as a Maximum Likelihood estimation problem, where the goal is to find the model structure (number of clusters) and parameters θ that best fits the data. A widely used solution to solve this problem is the EM algorithm. The algorithm takes as input the data and the desired number of clusters and outputs the model parameters and the posterior probabilities for each instance $\gamma_i(k)$ (the probability that an instance was generated by the k cluster). Note that the EM algorithm assumes that the number of clusters is known in advance, so that it does not directly tackle the problem of finding the model structure.

Evaluation of clustering results is again an application dependent problem. When the goal is to characterize the groups obtained, a strategy sometimes employed consists in defining a set of external characterization features that are not used in the learning process. For example, in customer segmentation for marketing purposes is usual to detect groups according to behavioral and demographic information and then complete the profile of these groups using business value characteristics such as profitability. In our experiments we follow a similar scheme using an external feature that indicates whether a student has pass or failed the course. We also make use of this feature in order to manually determine the most appropriate number of clusters from different EM runs. As a result, we obtain a profile for each cluster described both in terms of the input features and our notion of ‘profitability’ (student performance).

To help instructors interpret the clustering results in terms of the input features, we provide two additional pieces of information. First, we list all the features ordered by the degree of discrimination they provide between the different clusters. Additionally, we show a measure indicating how different the probability of each feature value in a given cluster is from the average probability in the full data. This measure of *lift*, commonly used to determine the interestingness of rules in association rule mining, reflects subsets of data inside a cluster that represent a behavior departing from the general tendency. The first information serves for the purpose of a general and comparative characterization of the groups while the second detects more particular behaviors derived from the segmentation obtained.

An examination of discriminant features for our first results indicated that the main characteristics used to form the clusters were the responses to the interests and skills surveys. While this clustering could make sense for some particular purpose, we found difficult to obtain some insight, especially because they exhibited little correlation with our profiling feature.

We concluded that the influence of survey related features was too high and modified the data including only two features indicating whether a student had answered each survey or not. With this modification we not only reduced the influence of these sorts of features, but also changed the perspective of the data. After some experiments,

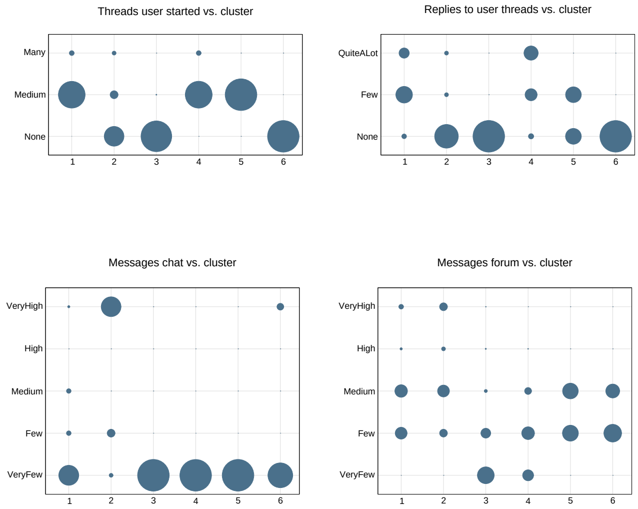

Figure 1. Cluster profiles.

we obtained a set of 6 clusters showing a non-perfect but still adequate correlation of the groups with the performance feature.

5 INTERPRETATION AND DEPLOYMENT

Figure 1 shows a graphical representation of four of the more discriminant properties obtained. The area of each circle represents the conditional probability of a value in a cluster. Note that there are some very frequent values that tend also to be frequent in most clusters, so the goal here is to look for similar patterns. For instance, the first graph shows the number of threads started and suggest that cluster 1 and 4 present a similar behavior, including students with higher initiative. Cluster 3 and 6 are groups that have never started a thread. By combining information of several features we can extract more particular profiles. For instance, cluster 1 and 4 can be further separated by considering the number of messages sent to the forum which shows that the first includes people with higher activity than the second. So being both clusters similar compared to the rest, members of cluster 4 seem to exhibit even a higher initiative since they start a lot of threads sending less messages to the forum.

Table 2 shows examples of group profiles obtained from the most discriminant features. Additionally, it includes interesting properties found in each cluster (differing from the average) but not covering all the members and the external profiling feature. They can be used, for example, to identify some effects derived from group behavior [9].

Students in cluster 1 are highly collaborative and appear to help their peers promoting discussion and, in some cases, creating new forums or publishing presentations. This cluster tends to represent our ideal collaboration profile.

In cluster 2, students tend to participate but at the same time their contributions do not generate too much interest. Since students appear to be motivated, a instructor could encourage this group to work harder in the course contents to better exploit their interactions.

Students in cluster 3 exhibit a passive behavior. They could be representative of free-riders, learners that do not know how to work in the group or simply lack of motivation for social work. Additional information on passive interactions (e.g. reading messages) would help to determine whether they are taking profit of their peer contributions or just worked alone. Depending on the interpretation, an instructor could encourage this group to work harder or provide clues about how to contribute, among other actions.

Cluster 4 group students that promote discussion but do not interact too much. In addition, they tend to fail the course. An explanation could be that their contributions were mainly off-topics of the course.

The group in cluster 5 exhibits an average interaction, with some of them having high external participation. Their good performance could be an indicator of a sort of gangling up effect, where they try to perform tasks as fast as possible. Instructors could convince these students for trying to help and support their peers.

Finally, cluster 6 are composed by students exhibiting an average degree of interaction, although some make extensive use of the chat. They might be receptive to further motivation, as suggested by some of them making use of a seldom used service like the file storage area or being aware of the course through alerts.

Results do not allow to conclude that collaboration is correlated with performance indicating that the effort of the instructors had limited impact in this regard. Results in clusters 2 and 3 suggest to try to further divide these groups in order to get additional insight.

In intermediate steps of the course, we could have used these profiles to help the tutor in creating groups of students to perform collaborative activities. Moreover, they can also exploited to assign roles into groups. For instance, members of cluster 1 could be a choice for moderators. Similarly, it can be taken into account that if cluster 3 represent free-riders, it may be counterproductive to join several of their members with a member of cluster 1, since this can promote the sucker effect. Obviously, these decisions would depend on the particular pedagogical strategy to be applied.

6 CONCLUDING REMARKS

The presented results, even preliminary, confirm the utility of data mining techniques to support evaluation of collaborative activities in

Cluster (%)	**Discriminant**	**Interesting**	**External**
1 (0.10)	High initiative, promote discussions, high participation in forum	Create forum, publish presentations, participate in external forums	Pass (0.77/0.23)
2 (0.07)	Medium/low initiative, don't promote discussions, high participation in forum, high participation in chat	Participate in external chats, create bookmarks, high number of sessions	Mixed (0.49/0.51)
3 (0.41)	Low initiative, low participation in forum and chat		Mixed (0.57/0.43)
4 (0.32)	High initiative, promote discussions, low participation in forum, low participation in chat		Fail (0.25/0.75)
5 (0.03)	Average in all areas	High number of visits to static pages, activate external alerts, participate in external forums, activate spam	Pass (1.0/0.0)
6 (0.07)	Low initiative, average participation in forum, extreme (low/high) participation in chat	Activate alerts, use file storage area	Pass (0.89/0.11)

Table 2. Student profiles including discriminant features, interesting features and an external profiling feature (pass/fail).

virtual communities. Particularly interesting is that the use of more elaborated features resulted in meaningful pattern descriptions. The subjective interpretation of these features made sense in the case studied but they may be only a rough approximation for other settings. We do not claim to be able to develop a universal solution, rather we advocate for deriving a large set of default elaborated features that may suggest more subjective, high-level interpretations and let the course instructors decide. Deriving new features taking into account knowledge about the goal is very common in business data mining and a well-known mean of improving the results [2].

A research direction to solve the previous problem is to define standard data models to represent collaborative interaction [7]. The preprocessing step will then consist in mapping original data to a higher-level, analysis oriented representation. Again, we can see a conceptual connection with business oriented data mining by considering data warehouses, specialized databases aimed to provide a business perspective of the data more suitable for analysis purposes.

Our discussion is also related to a recent proposal in [4] that proposes a classification of analysis methods for CSCL systems. Although our proposal could be considered a domain-independent approach, we are still studying this framework to see where data mining models fit.

A lesson learned from the analysis of this type of data is that data collection needs to be carefully designed and tuned to include all the possibly useful information. We take the risk of gathering too much data, but this is less of a problem that missing information, since data mining methods help to decide about which pieces are really relevant. Since we gathered the data using default or designed for other purposes tracking, we realized that some additional information could have improved our results, such as passive interactions. Further research and empirical evidence is needed in order to be more selective on the type of data needed and to relate this selection to particular pedagogical settings or goals.

Finally, focusing on clustering methods, we see at least two additional research opportunities. First, we can build clusters from low-level features to provide some guide to instructors about how higher level features can be derived for further analysis. For example, we could use the profiles for cluster 6 in the previous section to define a feature 'selective collaboration' for students that do not start threads, have average participation in forum but activate alerts or use the file storage area. A larger number of clusters would be probably more useful for this task. Secondly, clustering can be also directly applied to more elaborated data obtained in semi-structured workspaces, so that patterns can be automatically obtained instead of manually exploring individual or global reports.

We think that the data mining cycle, widely used for modeling business problems, fits very well into the recent line of research characterizing and classifying analysis methods for CSCL systems. We see a lot of promising research directions combining aspects from both views. Data mining can be a valuable source for data processing and model building techniques. In turn, CSCL research can provide methods to represent and integrate richer domain knowledge which, in fact, is still an open problem in data mining research.

REFERENCES

[1] B. Barros and M. F. Verdejo, 'Analysing student interaction processes in order to improve collaboration: the degree approach', *International Journal of Artificial Intelligence in Education*, **11**, 221–241, (2000).

[2] M. J. A. Berry and G. S. Linoff, *Mastering data mining. The art and science of customer relationship management*, Wiley & Sons, 2000.

[3] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'From data mining to knowledge discovery in databases', *AI Magazine*, **17**, 37–54, (1996).

[4] K. Gaßner, M. Jansen, A. Harrer, K. Herrmann, and H. U. Hoppe, 'Analysis methods for collaborative models and activities', in *Proceedings of the Conference on Computer Support for Collaborative Learning*, Bergen, Norway, (2003).

[5] P. Jermann, A. Soller, and M. Muehlenbrock, 'From mirroring to guiding: A review of state of the art technology for supporting collaborative learning', in *Proceedings of First European Conference on Computer-Supported Collaborative Learning*, ed., Kai Hakkarainen Pierre Dillenbourg, Anneke Eurolings, pp. 324–331, Mastrich, The Netherlands, (2001).

[6] A. Martínez, *Método y modelo para el apoyo computacional a la evaluación en CSCL*, Ph.D. dissertation, Faculty of Computer Science, Universidad de Valladolid, Spain, 2003.

[7] A. Martínez, Y. Dimitriadis, and P. de la Fuente, 'Towards an XML-based model for the representation of collaborative action', in *Proceedings of the Conference on Computer Support for Collaborative Learning*, Bergen, Norway, (2003).

[8] M. Meila and D. Heckerman, 'An experimental comparison of model-based clustering methods', *Machine Learning*, **42**(1/2), 9–29, (2001).

[9] G. Salomon and T. Globerson, 'When teams do not function the way they ought to', *International Journal of Education Research*, **13**(1), 89–100, (1989).