A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks

Antonio R. Anaya, Jesús G. Boticario
{arodriguez, jgb}@dia.uned.es
E.T.S.I.I. - UNED C/Juan del Rosal, 16, Ciudad Universitaria, 28040 Madrid (Spain)

Abstract. Data mining methods are successful in educational environments to discover new knowledge or learner skills or features. Unfortunately, they have not been used in depth with collaboration. We have developed a scalable data mining method, whose objective is to infer information on the collaboration during the collaboration process in a domain-independent way and to improve collaboration process management and learning in an open collaborative educational web environment. Thus, we used statistical indicators of learner's interactions in forums as the data source and a clustering algorithm to classify the data according to learner's collaboration. We showed the information on learner's collaboration to the tutor and learners to help them with collaboration process management. The experimental results support this method.

1 Introduction

Collaborative learning is a useful strategy to solve the lack of social interaction in most e-learning environments. However, the collaboration process has not been researched in depth [19]. We propose a data mining approach to reveal learner`s collaboration in open collaboration frameworks. We hypothesize that showing information on the collaboration process improves management and teaching in an open collaboration-learning environment.

The educational context of our research is suitable for e-learning and collaborative learning because of the nature of students at UNED (The National Distance Learning University in Spain). These students are mainly adults with responsibilities other than learning. For this reason UNED's students cannot be forced to collaborate in a typical CSCL (Computer Support Collaborative Learning) environment because of the time restrictions of these environments [8]. However, the collaboration learning is very suitable in the UNED's educational context due to the isolation of these students. We solved the problems by providing learners with an open collaborative learning experience, where students could manage their own collaborative learning process. We designed a long-term collaborative learning experience with fourth-year Artificial Intelligence (AI) and Engineering Based Knowledge students. This experience consisted of two main phases within a step-wise approach: the first phase was 3 weeks and the second phase was 10 weeks. It was enough time for students to complete the collaborative work and manage their collaborative process.

Although collaboration environments have been researched, some works have focused on monitoring learner interaction and showing this to learners [10, 5], but they have not used any inferring method to reveal learner collaboration. Others have concentrated on inferring information about the collaboration process [15], but the method used is not domain independent. We argue the need for inferring methods and the domain independence of these methods to be able to transfer the approach to other collaborative learning environments.

Given our educational context and our proposed open collaborative learning environment, we needed to simplify and reduce conceptual problems in order to improve collaboration process management and transfer the ideas to other environments. We achieved this by developing an inferring method that aimed to: 1) reveal learner's collaboration, 2) be domain independent, and 3) offer the information immediately after the process had finished. We applied the proposed approach to help students and tutors to manage the collaboration process by providing information on learner's interaction and learner's collaboration levels.

We covered the objective by using the statistics of interactions in forums as the data source and a clustering algorithm as the inferring method. Forums are a very common service in a collaborative learning environment and the statistics from forums can be obtained just after the interaction has happened. Since the statistics from forums do not give any semantic information, they are domain independent. [6] researched the forums messages in an educational environment and they concluded that the forums analysis can reveal the collaboration of the learners and an analysis by data mining is advisable. The statistical indicators were selected in relation to the learners' activity, initiative, regularity and promoting team-work [17]. That show, the method can be automated and the results are ready before the collaboration experience has finished. Those results provide information on the collaboration process to the tutor of the collaborative environment so that the tutor improved the teaching. Moreover, the same idea could be applied to learners. Thus we showed learner's collaboration levels to both the tutor and learners.

During the academic years 2006-07 and 2007-08 we researched the inferring method to reveal learner's collaboration [1]. During the collaborative learning experience of the year 2008-09 we showed learners the results of the inferring method. We concluded that the inferring method reveals learner's collaboration (more collaborative learners are more active and their activity encourages others to be more active) and the inferred representative collaboration indicators can be measured automatically.

A short overview of methods already used in evaluating the collaboration process is given below. We describe the collaborative learning experience and the inferring method. Next we show the results obtained after applying the inferring method and we explain in detail how the inferring information was shown to learners. Finally, we conclude with the discussion and future works.

2 Related Works

[16] said that the knowledge extraction process is divided into three phases: pre-processing, data mining and post-processing. In this section we describe research works that focus on collaboration process analysis and we explain the data acquisition method, the inferring or data mining method and the use of the results.

There have been various experiments to measure or identify the collaboration that takes place between system users. These experiments can be classified firstly the data acquisition method. We can identify three methods: 1) Qualitative [12]: where participants or experts are asked to evaluate the activities of the participants. 2) Quantitative [20, 15, 9, 3], which collects statistical information on the activities of the participants. 3) Mixed [4, 5, 10, 14]: where both methods are used simultaneously.

When we talk about data mining, these systems can be characterized by the inferring method used to derive the value of certain features, such as the collaboration that has occurred or is occurring. The methods may include: a) analysis by an expert [12], b) comparison with a pre-existing model using machine learning methods [15], c) Different statistical techniques [7] or machine learning, such as clustering [20, 14], fuzzy logic [15], sequential pattern mining [14], and d) the systems can be characterized even by not using any inference system [3, 4, 5, 10].

Finally, the systems can be characterized by the data post-processing method, or by what they do with the results. According to [18], CSCL systems, and in this case the systems that we are analyzing can be characterized by what they do with the results: I) monitoring tools that automatically collect data from students on the interaction, and they show this information [12, 3, 4, 5, 10], II) metacognitive tools that show the information inferred in the mining process, as well as interactions [15, 20, 14], and III) guidance system that proposes remedial measures to help the student, once the right information has been inferred. [6] proved the analysis of the forums by data mining and text mining techniques provide with meaningful feedback about student's performance and a view of the historical progress of a community of learners. We have followed the ideas but in a domain independence way.

We propose the data mining method, whose acquisition method is quantitative because the data source is statistical indicators of learner interaction in forums. As an inferring method we use the clustering algorithm, whose objective is to classify learners according to their collaboration, which is disclosed from learner's interaction. Finally, the proposed data mining method provides learners and tutors with metacognitive information on collaboration. The proposed data mining method is a metacognitive tool, which covers the research objectives.

3 Collaboration experience

We offered a collaborative learning experience to our learners with the objective of solving the common problems of distance learning and e-learning [10]. This research was carried out at UNED, where students are not typical university students. These students are mainly adults with responsibilities other than learning, who fit into the Lifelong Learning Paradigm, which supports the idea that learning should occur throughout a person's lifetime, enabling the integration of education and work in a continuous process. This impacts on the time that students can use to learn, study or take part in a typical CSCL system.

Figure 1 shows the schema of the collaboration learning experience. It was offered at the beginning of the academic years 2006-07, 2007-08 and 2008-09, lasted around 3 months, and was divided into two phases. The 1st phase lasted 3 weeks and learners had to answer an initial questionnaire and do a mandatory task. The 2nd phase grouped learners who had done the mandatory task into three-member teams, and the teams had to follow 6 tasks. Figure 1 shows the number of learners that started the 1st phase and finished it, and finished the 2nd phase. The Figure 1 show the schema of the collaborative learning experience, where there were two phases and some tasks had to be done in a sequential order.

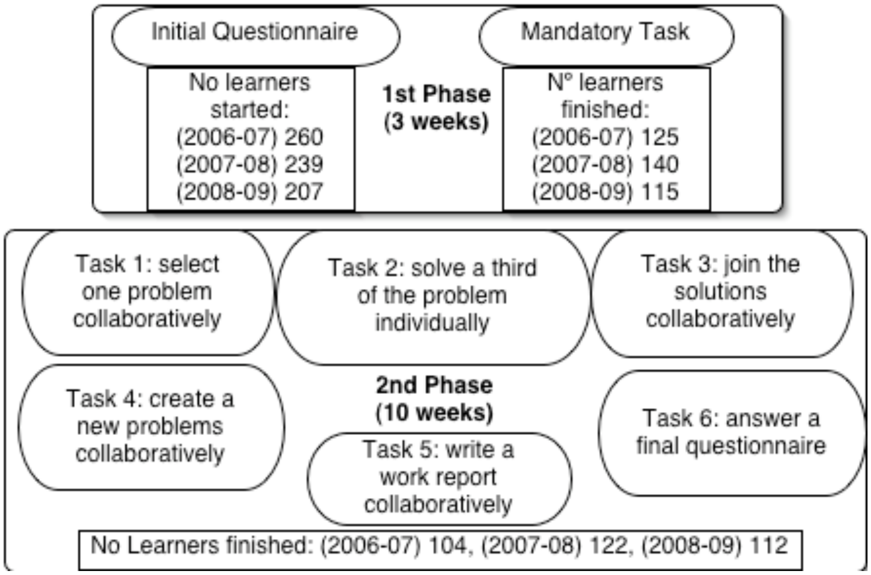
Initial Questionnaire
Mandatory Task
No learners started: (2006-07) 260 (2007-08) 239 (2008-09) 207
1st Phase (3 weeks)
N° learners finished: (2006-07) 125 (2007-08) 140 (2008-09) 115
Task 1: select one problem collaboratively
Task 2: solve a third of the problem individually
Task 3: join the solutions collaboratively
Task 4: create a new problems collaboratively
2nd Phase (10 weeks)
Task 5: write a work report collaboratively
Task 6: answer a final questionnaire
No Learners finished: (2006-07) 104, (2007-08) 122, (2008-09) 112

Figure 1. Schema of the collaborative learning experience

We provided a learning platform dotLRN (http://dotlrn.org/), which supports all learning experience activities, provides communications services such as forums, and stores all the interactions that take place on the platform in a relation database. During the 1st phase a general virtual environment was opened for all learners of the subject with common services (FAQs, news, surveys, calendar and forums). During the 2nd phase virtual spaces for each three-member team were opened, where the teams could perform the tasks. The specific virtual spaces include documents, surveys, news, a task manager and forums.

4 Method

We developed the inferring method with these objectives in mind: 1) the method should obtain information on learner's collaboration; 2) the method should be domain independent; 3) the method should provide information on collaboration before the collaboration process finished. Thus, it is possible reusing and applying the approach in other e-learning environments. We were looking approaches that could be applied to others at UNED, where there are 4000 curses and over 190000 students enrolled.

The learners of the collaborative learning experience were encouraged to use the forums on the dotLRN platform as the main communication media. The platform stores the forum messages giving information on what thread the messages are in and what message the message has replied to. We focused on forum interactions, because they are a very common service in a collaborative learning environment and the statistics from forums can be obtained just after the interaction has happened and data mining analysis is possible with these indicators [6, 8]. Since the statistics from forums do not give any semantic information, they are domain independent.

In line with the objectives explained above, we used statistical indicators of learner interaction in forums as a data source. According to [17], the features of collaborative learners in these environments are: activity, initiative, regularity and promoting team-work. We proposed these attributes as indicators of the above features: number of threads or conversations that the learner started (num_thrd), and their average, square variance and the number of threads divided by their variance; the number of messages sent (num_msg), and their average, square variance and the number of messages divided by their variance; the number of replies in the thread started by the user (num_reply_thrd), and divided by the number of user threads; the number of replies to messages sent by the user (num_reply_msg), and divided by the number of user messages. The number of threads started and their associated indicators are related to learner initiative. The square variance of the number of threads is related to the regularity of the initiative. The number of messages sent and their associated indicators are related to learner activity and regularity of activity. The number of replies to messages sent and their associated indicators are related to the activity caused by the learner.

We built datasets with the above statistical indicators from every year (2006-07, 2007-08 and 2008-09). The characteristics of the datasets were: Dataset-06-07, 117 instances; Dataset-07-08, 122 instances; Dataset-08-09, 112 instances. Every instance is the statistical indicators of the interactions of one learner. We focused our research on the collaborative period, which started at the end of November and finished at the end of January. We collected the values of these statistical indicators in datasets during the whole collaborative period.

We used a clustering algorithm as the data mining method. We used a clustering method because it classifies data collection without help from any expert, which delays the inferring process. We employed the EM clustering algorithm because of its good results when the method is applied in the learning environment to reveal collaboration. [20, 14, 13].

We obtained a classification of the instances with the EM clustering algorithm. We used the WEKA data mining software [21] and the EM clustering algorithm [7]. We checked the relation of the classification obtained with collaboration.

We needed to know student collaboration from another source to be able to compare their results and validate the approach as a collaborative inferring method. For this reason an expert identified student collaboration in the experiences. The expert read all the forum messages and labeled students according to their collaboration levels. Thus, we obtained a list of most of the students labeled according to their collaboration level. The expert used a scale of 8 values (1, low collaboration level; 9, high collaboration level).

Finally, the method finished by comparing the clustering classification of the learners with the labeled list of learner's collaboration levels. The objective was to measure the average collaboration level of each cluster and to realize that the average collaboration level is different in each cluster.

5 Results

We have conducted this research during the last three years. In 2006-07 and 2007-08 we focused on the aforementioned inferring method in order to prove the usefulness of the method as a collaboration inferring method. During 2008-09 we applied the method to improve collaborative process management and learning. We proved that the clusters obtained from statistical indicators were related to learner collaboration in the last two years [1] and the data for 2008-09 support these conclusions.

We classified the learners into 3 clusters, because the meaning of the classification is easier to understand in relation to collaboration. One cluster represents the low collaboration level, another cluster the medium collaboration level and the third cluster the high collaboration level. Then we run the clustering algorithm EM to obtain 3 cluster and we supplied with the datasets of every year (D-06-07, D-07-08 and D-08-09). These datasets collected the above statistical indicators for every learner.

First of all, we note that the cluster algorithm classifies according to the interaction. One cluster (cluster-0 in the next table) collects learners with low interaction (low values in the statistical indicators), another (cluster-1) collects learners with a medium level of interaction, and the third (cluster-2) collects learners with high interaction (high values of statistical indicators). Then we measured the average collaboration level in each cluster (column "Level" of the next table).

Table 1. Cluster collaboration level average

Dataset	Cluster-0			Cluster-1			Cluster-2		
	N_msg	N_reply_msg	Level	N_msg	N_reply_msg	Level	N_msg	N_reply_msg	Level
D-06-07	17.12	10.65	4.38	32.46	26.92	6.15	46.06	38.71	6.74
D-07-08	8.86	6.03	4.79	22.45	17.63	5.61	44.78	39.38	6.11
D-08-09	14.05	11.10	5.14	33.55	26.61	5.75	48.26	44.89	6.74

Table 1 shows the average of the statistical indicator "num_msg" (number of messages sent to the forums), "num_reply_msg" (number of replies to the messages sent to the forums), and the average collaboration level (Level), which was supplied by the expert, in every cluster. The table shows just two statistical indicators because they define the clusters better, although the clustering algorithm EM used datasets with the 12 statistical indicators, which were explained above.

We concluded that the relation between collaboration (collaboration level supplied by the expert) and the clusters, and the statistical indicators is clear. Therefore, the most active learners (cluster-2), i.e., who sent more messages and whose statistical indicator "num_msg" is higher, and who caused more activity (statistical indicator "num_reply_msg" is higher) are the most collaborative learners. From this we can label learners according to their collaboration. Clusters-0 learners are labeled with low collaboration level, cluster-1 learners are labeled with medium collaboration level, and cluster-2 learners are labeled with high collaboration level. Considering the coverage of the evaluations performed over three consecutive academic years and the number of students involved, we can conclude that the relation between the collaboration level and the inferred representative collaboration indicators can be measured automatically, which was done this 2008-09.

6 Result Management

The year 2008-09 we used this method and learner collaboration levels were calculated during the collaborative period. The objective was not to calculate the exactly collaboration level. We argue that calculating the exact value of one variable in an environment, which is in imperfect scientific conditions, is very complicated. The method used offers rough information on the collaboration level, which can be used to improve learning.

We thought that we could show the collaboration level to the tutor of the collaborative environment so that the tutor improved the teaching. The same idea, however, could be applied to learners. Thus we showed learner's collaboration levels to the tutor and learners.

We prepared different ways of showing the information to learners.

- Statistical indicator portlet. We prepared a tool displaying the value of only 4 statistical indicators (num_thrd, num_msg, num_reply_thrd and num_reply_msg) of every week during the collaboration period. The objective was to give information on the interaction during the collaborative process to team-members.

- Collaboration level portlet. We proved that our data mining method reveals the rough learner collaboration level. This tool displays the collaboration level of team-members and the information was updated every week until the end of the collaboration process. The objective was to give information on the collaboration behavior of team-members.

We offered these tools to 2008-09 students. The statistical indicator portlet was offered to 6 teams (18 learners), the collaboration level portlet was offered to 8 teams (24 learners), and both portlets were offered to 6 teams (18 learners). The collaborative learning experience finished, but the academic year has not finished. We are currently analyzing learners' answers to an opinion questionnaire and the collaboration learning experience results to prove the usefulness of the portlets. We offered these questionnaires to teams who had used some tool. The results are explained in the next table.

Table 2. Evaluation of tools

Tools	No. of learners who could use the tool	No. of answers	Average rank [0, 5]
Statistical indicators	18	9	3.33
Collaboration level	24	13	3
Statistical indicators and collaboration level	18	12	3.08

Half of the learners or more, to whom some tool was offered, answered the questionnaire and they had to rank the tools between 5 (highest value) and 0 (lowest). The average rank of every tool is not really high but it is always over half values (2.5). The results are positive but the poor number of answers means that we should be cautions on their analysis. To improve the analysis of the questionnaire we are comparing the above results with the marks and the collaboration period evaluation by the tutor. The aforementioned questionnaire will be contrasted with students' marks from tutors' evaluations and final exams. The latter will be available next June.

7 Conclusion and Future Work

In this paper we have proposed a data mining approach to improve teaching and learning awareness on collaboration features in open collaborative learning frameworks. It infers learner collaboration levels and shows this information to tutors and learners. We thought that the data mining method covers the objective needed to improve the collaboration process. The objectives are: obtaining information on learner collaboration just after collaboration interactions have finished and guarantee domain independency. These objectives guarantee the data mining method can apply to others.

This research focused on obtaining information on the collaboration process using statistical indicators of learner interaction in forums, machine learning technology as the inferring method, and showing the inferred information to tutors as the approach to improve the collaboration process. We have proposed statistical indicators, which are related to the activity: initiative, regularity of the learners and the activity caused by the learners. We think the above features explain the collaborative work [17]. An EM clustering algorithm classified the learner statistical indicators and learner collaboration levels, which were provided by an expert, were used to validate the clustering classification as a collaboration level classification. This research took place over three academic years 2006-07, 2007-08 and 2008-09, and more than 100 students took part in the collaborative learning experience each year (125 in 2006-07, 140 in 2007-08 and 115 in 2008-09). During 2006-07 and 2007-08 the research focused on the inferring method [1] and this 2008-09 the results inferred were shown to learners and their usefulness measured.

The results have proved that the data mining method could reveal representative collaboration indicators and help learners to improve collaboration learning management.

We have proved the clustering approach infers information on the learners' collaboration, but we do not have any empirical conclusion claim that the clustering method is better than other machine learning methods, which can adapt itself to the problem. To clarify this issue we are carrying out parallel research where the inferring method relies on decision tree algorithms [2]. We are currently collecting results from the datasets so that we can subsequently compare the new results from the application of decision tree algorithms with the results reported in this paper. Another open issue is evaluating the tools offered. To date the evaluation has given satisfactions, but the tools could be improved. However, we must be cautions and wait until the results from the opinion questionnaire and the results from the exams and collaboration experience evaluation by the tutor are compared and analyzed.

References

[1] Anaya, A.R., Boticario, J.G. Clustering Learners according to their Collaboration. 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009).

[2] Berikov, V., Litvinenko, A. Methods for statistical data analysis with decision trees. Novosibirsk, Sobolev Institute of Mathematics, (2003)

[3] Bratitis, T., Dimitracopoulou, A., Martínez-Monés, A., Marcos-García, J.A., Dimitriadis, Y. Supporting members of a learning community using interaction analysis tools: the example of the Kaleidoscope NoE scientific network Proceedings of the IEEE International Conference on Advanced Learning Technologies, ICALT 2008, 809-813, Santander, Spain, July 2008.

[4] Collazos, C.A., Guerrero, L.A., Pino, J.A., Renzi, S., Klobas, J., Ortega, M., Redondo, M.A., Bravo, C. Evaluating Collaborative Learning Processes using System-based Measurement. *Educational Technology & Society*, 10 (3), 257-274.

[5] Daradoumis, T., Martínez-Mónes, A., Xhafa, F. A Layered Framework for Evaluating OnLine Collaborative Learning Interactions". International Journal of Human-Computer Studies, Volume 64 , Issue 7 (July 2006), Pages 622-635

[6] Dringus, L.P., Ellis, E. Using data mining as a strategy for assessing asynchronous discussion forums. Computers & Education, 45(2005), 140-160.

[7] Gama, J., Gaber, M.M. (Eds), Learning from Data Streams: Processing Techniques in Sensor Networks, a book published by Springer Verlag, (2007)

[8] Gaudioso, E., Santos, O.C., Rodríguez, A., Boticario, J.G. A Proposal for Modelling a Collaborative Task in a Web-Based Learning Environment. 9th International Conference on User Modeling (UM'03). Workshop 'User and Group models for web-based adaptive collaborative environments'. Johnstown, Pennsylvania (United States)

[9] Hong, W. Spinning Your Course Into A Web Classroom - Advantages And Challenges. International Conference on Engineering Education August 6 – 10, 2001 Oslo, Norway

[10] Martínez, A., Dimitriadis, Y., Gómez, E., Jorrín, I., Rubia, B., Marcos, J.A. Studying participation networks in collaboration using mixed methods. International Journal of Computer-Supported Collaborative Learning. Volume 1, Number 3 / September 2006. 383-408

[11] Martínez, R., Bosch M., Herrero, M.M., Nuño, A.S. Psychopedagogical components and processes in e-learning. Lessons from an unsuccessful on-line course. Computers in Human Behavior 23, 146–161.

[12] Meier, A., Spada, H., Rummel, N. A rating scheme for assessing the quality of computer-supported collaboration processes. Computer-Supported Collaborative Learning (2) (2006) 63–86

[13] Meilâ, M., Heckerman, D. An Experimental Comparison of Model-Based Clustering Methods. Machine Learning, 42, 9–29, 2001

[14] Perera, D., Kay, J., Yacef, K., Koprinska, I. Mining learners' traces from an online collaboration tool. Workshop of Educational Data Mining (AIED'07).

[15] Redondo, M.A., Bravo, C., Bravo, J., Ortega, M. Applying Fuzzy Logic to Analyze Collaborative Learning Experiences in an e-Learning Environment.USDLA Journal. (United States Distance Learning Association).17.2, 19-28.

[16] Romero, C., Ventura. S. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications 33 (2007) 135–146

[17] Santos, O.C., Rodríguez, A., Gaudioso, E., Boticario, J.G. Helping the tutor to manage a collaborative task in a web-based learning environment. In: AIED 2003: Supplementary Proceedings. (2003) 153–162

[18] Soller, A., Martinez, A., Jermann, P., Muehlenbrock, M. From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. International Journal of Artificial Intelligence in Education, vol. 15, 2005, p. 261-290

[19] Strijbos, J-W., Fischer, F. Methodological challenges for collaborative learning research. Learning and Instruction 17 (2007)

[20] Talavera, L., Gaudioso, E. Mining Student Data To Characterize Similar Behavior Groups In Unstructured Collaboration Spaces. In: Proceedings of the Workshop on Artificial Intelligence in CSCL. 16th European Conference on Artificial Intelligence, (ECAI 2004), Valencia, Spain. 17–23. (2004)

[21] Witten, I. H., Frank, E. Data Mining. Morgan Kaufmann, June (2005).