TESIS DOCTORAL

Prospección de la colaboración utilizando herramientas de minería de datos en ambiente abiertos de aprendizaje colaborativo con el objetivo de mejorar la gestión del proceso de colaboración

Antonio Rodríguez Anaya

Licenciado en Ciencias Físicas

por la Universidad Complutense de Madrid

Dpto. de Inteligencia Artificial

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

UNED,

2009

Departamento de Inteligencia Artificial,

Escuela Técnica Superior de Ingeniería Informática.

Prospección de la colaboración utilizando herramientas de minería de datos en ambiente abiertos de aprendizaje colaborativo con el objetivo de mejorar la gestión del proceso de colaboración

Antonio Rodríguez Anaya

Licenciado en Ciencias Físicas

por la Universidad Complutense de Madrid

Jesús González Boticario

1.1 Resumen y conclusiones de la tesis de Antonio Rodríguez Anaya

Summary

The Distance Education (DE) is an education model has a number of relevant features, such as breaking the barriers of time and place. The teacher can instruct the student without having to be both on the same site and at the same time. This advantage brings the disadvantage that the communication between them has to use other communication means less rich and intense than face to face. The great advantage of the ED is paradoxically also the main constraint (Garcia Aretio, 2009).

The motivation of this thesis has been the improvement of learning in an environment of DE over the Internet through the strategy of collaborative learning. Although the advantages seem obvious use of collaboration in distance education, as noted, a study of collaboration is necessary to try to ensure that collaborative learning is performed.

The design of web collaborative learning environments must primarily ensure collaborative learning that is the same or better than individual or competitive learning. According to Johnson & Johnson (1989), offering students in these environments a set of communication tools and a set of tasks to complete is not enough. These authors establish five conditions to make collaborative learning better than individual or competitive learning (Johnson & Johnson, 2004): (1) positive interdependence (everyone shares the goals), (2) individual accountability/personal responsibility (everyone is in charge of oneself), (3) promote interaction (mainly face to face interaction), (4) interpersonal and small group kills (everyone works effectively with each other and functions as part of a group), and (5) frequent and regular processing of the group's functioning to improve its effectiveness in the future should all be clearly perceived.

In e-learning environments, the use of collaboration strategy is highly recommendable to minimize disadvantages inherent to distance learning. In this context communication between students can be asynchronous, where students are free to perform communication acts when they want (Santos & Boticario, 2004b), i.e.

they control the collaboration process. As for managing collaborative experiences in these contexts, the tutor has a very complicated task if s/he is in charge of analyzing regular and frequent collaboration, especially if the courses are distant teaching and involve over a hundred students.

Given the previous conditions, the design of the collaborative learning environment should therefore use a method that analyses student collaboration regularly and frequently with little or no intervention by the tutor of the learning experience. Learner modeling techniques are appropriate in these contexts because they have been used to analyze the learners' behavior with the objectives to model the user to use this model so that the system or the learner improves the learning, in a educational environment (Brusilovsky & Millan, 2007). So that the modeling techniques are automatic or they can be used with little or no intervention by the tutor of the learning experience, we can require data mining methodology (Romero & Ventura, 2007) and use machine learning technology (Russell & Norvig, 1995).

Motivation and problem definition of this thesis have addressed the same objectives, in addition to the assumptions used for solving the problem defined. To carry out research have postulated a set of hypotheses:

1. it is possible to characterize the collaboration in open environments using technology instead of an expert analysis,
2. increasing control over the collaborative process encourages collaboration and improves the management process itself,
3. increasing control to be effective must be done while the process of collaboration is taking place, and
4. an analysis of collaboration without domain knowledge is possible and which allows repeated in other educational settings.

The hypotheses require evaluation of research in a real collaborative learning and establish a set of objectives. The overall objectives were:

1. characterize their students on their collaboration,
2. use a method of analysis to get results on a regular and frequent
3. characterization to provide students and assess their learning, and
4. offer a general solution to the work of this thesis are applicable in other collaborative learning environments.

Framework

The work of improving the collaborative learning modeling the learners' collaboration is a research, which can be divided in (Romero and Ventura, 2007; Kobsa, 2001; Soller et al., 2005): the data acquisition methods, the inference method, the model and the process that they apply to their results and model.

Data acquisition methods are: a) Qualitative (Meier et al., 2006), where the individuals participating in the research are directly questioned, or experts assess participants' activities; b) Quantitative (Talavera & Gaudioso, 2004; Redondo et al., 2003; Hong, 2001; Bratitsis et al., 2008), where statistical information on participants' activities is collected; and C) Mixed (Collazos et al., 2007; Daradoumis et al., 2006; Martínez et al., 2006; Perera el al., 2007), the use of both methods simultaneously.

Systems can be characterized according to the inference method used to deduce the value of certain characteristics, such as collaboration, which has occurred or is occurring. The methods may be: 1) an expert's analysis (Meier et al., 2006); 2) comparison with a pre-existing model (Redondo et al., 2003); 3) different statistical or interaction analysis techniques (Hong, 2001; Daradoumis et al., 2006; Martínez et al., 2006; Bratitsis et al., 2008) ML techniques, like decision trees, clustering, sequence pattern mining, etc., (Talavera & Gaudioso, 2004; Redondo et al., 2003; Perera el al., 2007); 4) even characterizing systems by not using any inference system (Collazos et al. 2007;).

After the inferring method and before of using the results of the method, some researches have model the learners' collaboration. Two main strategies have been identified. There has been research that has proposed a model for collaboration (Redondo et al., 2003; Baghaei & Mitrovic, 2007; Martínez et al., 2003; Vidou et al., 2006; Barros et al., 2001; Barros et al., 2002; Durán, 2006), while others have proposed a number of indicators related to collaboration (Collazos et al., 2007; Baldiris et al., 2007; Soller, 2001; Park & Hyun, 2006; Baeza-Yates & Pino, 2006; Meier et al., 2007; Kahrimanis et al., 2009; Bayón et al., 2007; Martínez el al., 2006; Daradoumis et al., 2006; Bratitsis et al., 2008). As many have noted (Park and Hyun, 2006; Strijbos and Fischer, 2007), analysis or modeling of the collaboration is a field still open and it needs a deep work to standardize the field and build methodologies.

It has been said that to ensure that collaborative learning takes place, it is necessary to analyze student collaboration regularly and frequently. In order to achieve this, an analysis method and a tool that uses the inferred information have to be planned. In collaborative environments these tools can be classified according to their function (Soller et al., 2005): I) Monitoring tools that collect data on student interaction automatically and show this information (Meier et al., 2007; Kahrimanis et al., 2009; Collazos et al., 2007; Daradoumis et al., 2006; Martínez et al., 2006; Bratitsis et al., 2008); II) Metacognitive tools that show, as well as the interactions, the information inferred from processing the collaboration analysis, i.e., they propose judgments (Redondo et al., 2003; Talavera & Gaudioso, 2004; Perera el al., 2007); III) Guide tools, which propose corrective measures to help the student, once the right information has been inferred (Baghaei & Mitrovic, 2007; Santos & Boticario, 2008).

In other words, there are tools (I) whose aim is to show the data on students, so the priority is to discover how these data are acquired, filter them and display them. For other tools (II) the priority is to infer useful information from the data collected, and others (III) guide or recommend students thanks to the results inferred.

A useful strategy on the educational field is the self-regulation. A self-regulation tool uses metacognitive techniques to discover what has been learnt (Hartman, 2001). It has been shown that helping students use their self-regulation skills improves their learning (Pintrich, 2000) and dividing self-regulation into three characteristics or tasks (Zimmerman, 1990) has even been proposed: i) self-observation (tracking of own activities), j) self-judgment (self-assessment of performance) and, k) self-reactions (reactions to performance results).

Moreover, (Stiffens, 2001) establishes that self-regulation is more than the regulation of own cognitive activities (metacognition), since it also involves motivational and emotional aspects. It is obvious, according to the "self-regulation handbook " (Beakers et al., 2000), that self-regulation is not only used in monitoring learning processes of oneself, but it also plays an important role in managing social activities.

From what has been said about self-regulation it is deduced that a metacognitive tools fulfils the conditions to be considered as a self-regulation tool. A metacognitive tool monitors students, provides them with an independent judgment that may change

according to their reactions, so they can assess themselves and be aware of their reactions.

Studying the literature on the topic we have observed that two types of tools behave like self-regulation tools: 1) those that show the information to students, and 2) those that allow students to manage their own data. These second tools are known as scrutable (Kay, 1999).

Some research studies pose scrutable systems to improve learning. They have noticed improved understanding and control of learning in educational environments that use scrutability and open models (Bull & Gardner, 2009), which are student models that can be accessed and managed by the student (Bull & Kay, 2007). Scrutable systems are gaining interest in educational environments, as some research studies show (Verpoorten et al., 2009).

Critical Analysis

When establishing a framework for the improvement of collaborative learning, we have studied the literature that addresses the problem or any of its parts. With the above, the design of the collaborative learning environment must be considered analyzing the behavior of participants on a regular and frequent with less intervention of tutors, because of the context of the DE. The characterizations of the researches in the field can be divided into: how to obtain data (preprocessing), which method of analysis used (inference), what type of modeling is applied (modeling), what type of tool used (tool).

Due to the lack of methodology, standards and empirical studies (Strijbos & Fischer, 2007, Perera et al., 2007; Bratitsis & Dimitracopoulou, 2006), there are several open issues in the analysis of collaboration. The first is the data source. Due to the educational context and objectives of the investigation, it was thought that thing was used as a source of data, quantitative information about the interactions of students relating to collaboration. In this way promotes regular and frequent collection of data, and the lowest possible intervention of tutors and students.

Another open issue in the analysis of collaboration is the method of inference. Since the analysis should be performed regularly and frequently with the least

possible intervention of tutors or experts, it is necessary to use machine learning techniques.

The analysis of collaboration has been the main part of the modeling of the student, but not the only one. Modeling the student is a major challenge for the personalization of learning environments (McCalla et al., 2000). A very interesting strategy from the pedagogical point of view is the open model (Bull and Kay, 2008), where the result of modeling is shown to the student and where he can manage. Therefore the model obtained from the modeling process should be understandable by the student.

In collaborative environments thus information on the process and in the context of cooperation are of interest (Muehlenbrock, 2005). We conclude that useful information for the student to work is due to the context of collaboration, the process of collaboration, and inferred by the analysis as it provides judgments about student collaboration.

To close the cycle of improvement of collaborative learning, we have discussed various pedagogical tools used. We have focused on the benefits of self-regulatory tools in the educational context and are associated with collaboration, as the social processes involved in self-regulation (Boekaerts et al., 2000). Characteristics have been established self-regulatory tools and compared with metacognitive tools that can provide a collaborative environment. We have opted in this investigation by a self-regulatory metacognitive tool.

Following the thread of the models open to students, we have studied the tellers (Kay, 1999). In defining the searching has found affinity with the searching tools and metacognitive self-regulation. It has been reported the theoretical advantages of the teller to increase accountability and, therefore, motivation and active learning (Hummel et al., 2005, Burleson, 2005; Boticario and Gaudioso, 2000a).

In summary, we conclude that a collaborative learning environment via the web within the paradigm of the ED requires the system to perform analysis of collaboration on a regular and frequent. Through the analysis can be modeled to students. Once the modeling result, ie, the model can become a self-regulatory metacognitive tool with which to provide the student with a means of helping manage the collaborative process. This tool can add the benefits of teller to increase student responsibility in the entire collaborative process.

Research

This research was done at UNED (National University for Distance Education, Spain). UNED students are adults with responsibilities other than learning. Thus, it is difficult for them to participate in collaborative environments with the typical time restrictions (Santos et al., 2003), since they must control their learning processes (Gaudioso et al., 2003). In order to improve the learning of students on the fourth year Artificial Intelligence and Knowledge-based Engineering (AI-KE) subject at UNED Computer Engineering School, during the academic years 2006-07, 2007-08 and 2008-09 they were offered a long-term collaborative learning experience where collaboration was controlled and managed by the students. Over 100 students participated in each collaborative learning experience over the three consecutive years.

Given the research aims and the educational context, ML techniques are advisable to analyze student interaction. With this approach one difficulty is the lack of methodology and standards to analyze collaboration (Strijbos & Fisher, 2007), as well as not being able to identify beforehand the most appropriate ML technique for the problem. Therefore, an empirical study is necessary. This work presents two approaches to analyze collaboration regularly and frequently using different ML techniques; the results are analyzed and compared.

The two approaches use student interactions in forums as a source of data, like other related research studies (Bratitsis et al., 2008; Gómez-Sánchez et al. 2009), and their main difference is: 1) classifying students according to their collaboration using unsupervised classification techniques (clustering) (Anaya & Boticario, 2009a); 2) constructing collaboration metrics, using supervised classification techniques (decision tree algorithms), which assign each student with a collaboration value so that students can be compared. Using the statistics of forum interactions is justified because forums are a communication service widely used in e-learning environments, because they were the main means of communication in our collaborative learning experience, and because these statistics are an effective way of obtaining information on forum users, such as their level of activity or frequency of use (Dringus & Ellis, 2005).

Alter the collaboration analysis a model of collaboration has been built. The model structures information on collaboration and it is divided in: information on the collaboration context (personal data, academic data, work data, collaboration facilities data), information on the collaboration process (statistical indicators of learners' interactions in forums) and, evaluations on learners' collaboration (obtained by inferring method). The model uses the scrutability strategy and shows the data so that learners can understand and use to improve the collaboration learning (Bull & Kay, 2008).

Four tools were provided during the academic year 2008-09, because we wanted to research what type of information was the most useful to improve the collaboration process, and how to present it. Two presentation approaches were used. On the one hand, the simplest way for students to use and understand information (Barkley et al., 2004) and, on the other, the scrutable strategy was used.

Once the collaborative learning experience had finished, the tools were evaluated bearing in mind the students' opinions, the collaborative work rating and the learning rating. The results show that the information from the interaction analysis is useful to improve collaboration, but the most useful information is from the inference method. The results also showed that, in this educational context, the tools that only show or display the information collected or inferred help more than the tool that shows it scrutably.

Results

The results obtained with the data from the three collaborative learning experiences show that students can be classified according to their collaboration using clustering techniques, although their inferences are approximate.

They also show that collaboration metrics can be constructed to give a value to each student's collaboration. The collaboration results are also approximate in this second approach.

The research ended by comparing both approaches. Accordingly, two variables were proposed: the difference (Δ) of average values of the metrics between the different collaboration levels and error (E%) as the average percentage of variance for the metric value in each instance. Obviously, as the difference (Δ) is greater between

levels it is assumed that the approach could better distinguish student collaboration, and less error (E%) would make the result more certain. Comparative analysis has told us that the approach using clustering techniques provides inference with less accuracy than the metric approach.

During the 2008-09 academic year experience the interactions were analyzed and the inference method was done that provided information on student collaboration. Four tools were offered that provided the former information to check: 1) the type of information on the most useful collaboration, and 2) the type of strategy for presenting the information. Accordingly, a tool was designed as a web application that showed all the information on the collaboration scrutably (Kay, 1999), and three portlets were created, small windows in the platform, which showed information on the collaboration process (Portlet I), on student collaboration inference (Portlet II), and another that showed both types of information (Portlet III).

The results in this article indicate that: 1) the most useful information to improve collaborative learning is that inferred on student collaboration, i.e., student levels of collaboration, and 2) the students participating in the research made the most of that information that was shown in a more simple form, so the scrutable strategy, used in the web application, has not achieved the aims.

Conclusions

The research achievements are divided in: the process of analysis of collaboration, the proposed model of collaboration and the tools that have been proposed to improve the collaborative process.

1. The analysis of collaboration is divided into data acquisition and method of inference. To achieve the objectives defined for this part, data acquisition method must be quantitative to get the data regularly and frequently, evaluations of collaboration are necessary to ensure that collaborative learning occurs, machine learning techniques can be used to infer assessments collaboration. A method of analysis that will follow these arguments can be automated and has great potential to be transferred to other settings. In the thesis two approaches are described which have been validated in Chapter V. Therefore, an achievement of the thesis was: to find instead of one, two inferring methods of collaboration with which

assessments of the collaboration are obtained regularly and frequently, and they can be automated and transferred to other environments.

2. There is a lack of agreement and the lack of comparative studies in the collaborative modeling field (Bratitsis and Dimitracopoulou, 2006). We have followed a strategy aimed at expanding the learners control over their processes, and use is given directly to the user. This strategy is called the open model (Bull and Kay, 2008). We have followed this approach because the information structure in a model, which can be understood as an ontology (Mizoguchi, 2005), the learners uses the model, thus the model needs to be used and understood by the learner. These ideas provide pedagogical characteristics, since its objective is to improve learning, and reusability in other environments and contexts. Both features are appropriate to the objectives of the research. We have proposed a collaborative model divided into three parts to receive background information on the process and assessments of the collaboration. The achievement has been to create a model of collaboration to report on relevant aspects of the learners' collaboration, which is understandable and easily transferable to other environments.
3. To close the cycle of research in educational settings, we studied the possible corrective actions to be undertaken to ensure that collaborative learning takes place and also providing more control over the collaboration process. Studying the problem we have noted the benefits that a self-regulatory tool provides an educational and collaborative environments (Steffens, 2001). This idea is closely related to the open model (Bull and Kay, 2008) and metacognitive tools (Soller et al., 2005). We have suggested to students four self-regulatory tools in order to check the type of information more useful and how or tools. For these reasons it has developed a tool scrutable (Kay, 1999) that displayed all the collaborative model, and three other self-regulatory tools, one monitoring tool, which informs on the process of collaboration, and two that have already been metacognitive, which report the evaluation of the collaboration. An evaluation process was conducted. The achievements of this part are: check the advantages in working as a self-regulatory metacognitive tool in collaborative learning environments.

This thesis has as general achievement, which has proposed a methodology with the motivation to improve collaborative learning increasing control over the collaborative process. The collaboration has been analyzed and modeled the

collaboration, has proposed a model which has been used as the basis for a multitude of self-regulation that have been evaluated to see if targets are met. It has been shown that a relationship exists between the work (analysis model) with the increase of collaborative learning. We must say it has succeeded in achieving the improvement of collaborative learning, where students, tutors or experts have the least intervention possible, automation has been proposed that makes possible the transfer of the method to other environments, something that has not yet been tested.

Future works

The empirical study done in this research has to continue. Two ML techniques that were thought to be more appropriate for the problem were chosen. However, other ML techniques cannot be excluded, like Bagging. With this technique the statistical indicators most related to collaboration could be identified instead of using decision tree algorithms. From applying Bagging, other collaboration metrics could be constructed.

The results obtained are approximates. In both the approaches researched we have to continue to reduce the percentages of error and increase the difference between the levels of collaboration. The source of data used cannot be forgotten. Statistical indicators of students' active interactions in forums were used. The source of data could be increased by also using passive interactions (Burr & Spennemann, 2004), understood as the reading or visualisation of forum messages, so that they reinforce the collaboration related tasks. The use of some domain-related indicators can also be studied (Ravi et al., 2007; Lin et al., 2009). These indicators could add much information to the analysis, although we would have to consider that the analysis might then not be able to be transferred to other environments.

The results obtained indicate that collaboration analysis, as has been explained, provides useful information to the students on their own collaboration and that of their classmates, and students used it to improve collaboration. Learning has increased but to a less extent according to the students' subject exam results. This is a point open to research. It is very difficult to achieve results that unequivocally show the correlation between learning and access to the information provided, although the evidence obtained is encouraging. Other many dependent variables should be considered, like

for example the time devoted, which can depend on personal situations that were not controlled because of the inherent difficulty and so as not to disturb the students any further. It should be borne in mind that they are students following a distance education model. In future experiences this point will have to be clarified. Another point that this research leaves open is the usefulness of the scrutable strategy. The results were negative, but the limited use of the web application makes it impossible to pronounce judgment, other than that the tool in itself is not motivating. In future experiences the use of the web application is going to be encouraged improving usability integrating it into the dotLRN platform. Once it has been integrated and reasonably used by students, it will be possible to know whether in the educational domain posed scrutability provides students with greater responsibility and whether this improves the collaboration and learning.