Mining Techniques for Models of Collaborative Learning

Thereza P. P. Padilha[1,2], Leandro M. Almeida[1] and João B. M. Alves[2]
[1]Lutheran University of Brazil - ULBRA
Email: {thereza, leandro}@ulbra-to.br

[2]Federal University of Santa Catarina - UFSC
Email: {thereza, jbosco}@inf.ufsc.br

Abstract: Many student interactions in the collaborative learning process can be captured and stored in a database for future analysis. However, the precious information extraction in database is almost impossible without the use of mining techniques. In this paper we present a model of collaborative learning, individual and group, using data and text mining techniques. Our model allows us to extract relevant information about collaborative learning interactions at different levels of abstraction.

Keywords: collaborative learning, data mining, interaction data, text mining.

1 Introduction

The analysis of collaborative learning interactions is considered a key issue and powerful because it allows us to known and "understand" how learning evolution among students happens, for example. Several computational models of collaborative learning are found in literature, such as finite state machines (McManus & Aiken, 1995) and rule learning (Katz et al., 1999). Each one of these models has a different perspective. A review of some existing models can be found in (Soller & Lesgold, 2000). Before building models, it is necessary to identify the variables that are to be modeled. This is a difficult step because the specific variables that play an important role in this complex process are deeply entangled and, therefore hard to isolate in research (Pol, 2002). The next step towards the building of computational models is to analyze variables values. The analysis process is essential because in interactions' data (logfile) can be stored unnecessary, missing and redundant data. Thus, the use of mining techniques for processing large amount of logfile is indispensable (Martinez et al., 2002).

In this context, this paper presents a model of collaborative learning using data and text mining techniques. The model provides set of performance reports that allow us, for example, to compare a specific group with other groups or a member with the other group' members and verify the student actions historical. For this purpose, we use data mining (DM) techniques to compare the current state of interaction with the ideal state, and text mining (TM) techniques to identify and categorize contribution types in the dialogs. Our interest is to discuss related issues to models of collaborative learning and, mainly, the question "What compilation or abstraction methods are needed to construct a computational model from a logfile describing the group interaction?", described in the call for participation.

This paper is organized as follows. In section 2 we describe an overview data and text mining techniques. In section 3 we describe the functionality of the proposed model. In section 4 we present the conclusions and then comment some topics for discussion in the workshop.

2 Data and Text Mining - overview

DM is a technique that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations produce a particular enumeration of patterns (or models) over the data (Fayyad

et al., 1996). Data mining has been directed to search patterns from data set using methods such as neural networks, symbolic machine learning algorithms, probabilistic reasoning, etc. In the symbolic algorithms field, actually, there has been much interest in the semi-supervised learning, an intermediate type between supervised and unsupervised learning. In this context, learning refers to rules set, for instance. The semi-supervised learning has as main characteristic the incorporation of background knowledge through labeled examples in unlabeled data set for future learning (Bruce, 2001). Initially, a supervised learner is build using the labeled examples and then applies the trained learner on unlabeled data. There is not a pre-defined amount of labeled examples that should be inserted in database, however, if one database contains a high number of labeled examples more easy and correct will be its works.

The semi-supervised learning was chosen because of its flexibility and accuracy to use incorporated knowledge (ideal state), represented by labeled examples in the data set, and to classify the students' performance, represented by unlabeled examples, in collaborative process. For each realized classification, it is possible to know its accuracy level and the used patterns for definition of the value. Another reason is the ability to work with an undetermined amount of examples, but it is important to provide a minimum quantity of data.

TM is a technique that looking for regularities, patterns or trends in natural language text from unstructured or semi-structured texts (Tan, 1999). TM includes several text processing and classification tasks such as text categorization, clustering, summarization, information retrieval, etc. The text categorization, for example, is one task for labeling natural language texts with thematic categories from a predefined set. The categorization is realized via similarity measure assigning a Boolean value to each pair $(d_j, c_i) \in D \times C$, where D is a domain of documents and $C = \{c_1, c_2 \ldots, c_j\}$ is a set of predefined categories (Sebastiani, 2002). The categories are just symbolic labels. Categorization using Boolean model is simple because only verify, in D, if there is the presence of one or more words stored in C, for instance, to classify it.

In the model, we use the text categorization task to identify the student intentions such as task division, decision making and explanation among messages sent. For this identification, we need to build a set of predefined categories to evaluate the semantic. The principal advantage is to eliminate the dependence on users to provide their contribution types. However, for each application domain will require a specific set of category.

3 Mining Techniques to Model Students' Interactions

The proposed model is incorporated in a collaborative problem solving environment, implemented in Java, in which the collaboration is based on five steps that are: reality observation, key-points, theorization, hypothesis elaboration and reality fitting (Padilha, 2003). To facilitate the building model, two computational agents named awareness and collaborative were defined and will be described in more detail.

3.1 Awareness Agent

The awareness agent's goal is to capture, categorize and store several contribution types. The students' interactions (actions) for problem solving are stored in a MySQL database. We identified a set of quantitative and qualitative variables for modeling. The quantitative variables, basically, inform individual and group interactions number using communication tools or other resources available. On the other hand, the qualitative variables provide a social and cognitive aspect of interaction through actions performed by students. Table 1 presents a brief description of the variables used for modeling. The identification and categorization of the variables were realized with help out from several educators and psychologist.

Although awareness agent is implemented, we build a simple data set containing 40 examples about student interactions in the problem solving steps. Afterward, we added 10 labeled examples representing the ideal state. Thus, the data set consisted of 50 examples, in which 80% of unlabeled examples and 20% of labeled examples, and

11 attributes being 10 related with quantitative and qualitative variables and 1 to performance value. The quantitative variables values are numeric. The qualitative variables values are categorical that are: low, middle and high. These values are defined by observing and analyzing the group's conversation and actions in order to identify situations in which students effectively acquired knowledge. For this analysis, we determine a set of 5 predefined categories (organization, argumentation, information, request and motivation) to use the text categorization task. Each one of these categories is associated with one qualitative variable. The organization category supports the task division variable; argumentation and request support explanation variable; information supports decision making variable; and motivation supports involvement variable. Every category consists of, in average, 20 words.

To improve performance in the categorization and reduce search time, a selection process is realized for find adequate words. So, when a message is analyzed, "irrelevant" words (*stopwords*), such as articles and prepositions, are ignored. For example, the message *"How will we do the work?"* has a high probability for task division variable because it has the "work" private word belongs to organization category and the adverb "how". The implementation of the text categorization follows methods presented in (Vapnik, 1998).

3.2 Collaboration Agent

The collaboration agent works to build the model and produce some performance reports. First, the labeled examples are provided to a learning algorithm that generates a learner (patterns) analyzing its attributes' values. The patterns are represented as production rules, i.e. if-then format. Second, the unlabeled examples are submitted to rules that classify its performance. The performance value is numeric between 0 (bad) and 10 (excellent). The pattern discovery process is based in the Entropy and Gain Ratio methods. The Entropy is responsible to measure disorder level in attributes (Williamson, 2002). Entropy is high if there is a lot of disorder, otherwise it is low. Gain Ratio is responsible to obtain the amount of relevant information in a specific attribute (Quinlan, 1996).

Table 1: Variables for Modeling

Type	Name	Goal	Actions
Quantitive	Chat	To obtain the interactions number in synchronous discussion sessions in chat tool.	The computation of this variable occurs when the students send messages, use sentences openers, and access last discussions. Messages without content, with repetition of only one character or high number of symbols are not computed.
	Text Editor	To obtain the interactions number in the textual edition tool.	The computation of this variable occurs when the students create blocks in exist files, create a new file, read, write or change texts.
	Vote	To obtain the interactions number in vote tool, directed, mainly, to decide ways that the students should follow.	The computation of this variable occurs when the students create a new voting or vote in one current voting.
	Form	To obtain the number of participations in the form filling, needed for problem solving.	The computation of this variable occurs when the students forms filling for problem solving.
	Repository	To obtain the interactions number in the repository.	The computation of this variable occurs when the students make document upload or download.
Qual	Decision Making	To identify the degree of decision making for problem solving.	The computation of this variable occurs through a semantic analysis in messages sent.
	Task Division	To identify the degree of organization among students to divide theirs tasks.	The computation of this variable occurs through a semantic analysis in messages sent.

i t a t i v e	Regularity	To identify the degree of access to environment and tools during problem solving as a whole. The regularity reflects the interest and responsibility with other group members.	The computation of this variable occurs when the students access and use environment, communication tools, and other resources available.
	Explanation	To identify the degree of explanation/argumentation by students for discussing related topics with problem.	The computation of this variable occurs through a semantic analysis in messages sent.
	Involvement	To identify the degree of interactivity among group members. The involvement is a variable very important to indicate the presence of communication among students.	The computation of this variable occurs through an analysis of interactions realized among students using communication tools and other available resources. However, if a student has high value in use of the chat tool but not in vote tool, then he/she will be considered medium involvement. Moreover, a semantic analysis occurs in messages sent.

The collaboration agent provides a set of reports presenting an overview of the student performance. The reports have different levels of abstraction: comparing a specific group with other groups, comparing a member with other group' members and analyzing of student actions historical. The Figure 1(A) presents, graphically, the group performance report during the solving of four problems. For each existing group, there is a specific line format that it is possible to observe its performance and verify the global performance (available in the below legend). The group B, for example, had the bad performance as can be seen in its line format. The student performance report is very similar to group performance report. In this case, it is possible to verify the performance of each student in a group. The Figure 1(B) shows the student performance report of the group A.

The Figure 1(C) illustrates, in more details, all students' actions in key-points step of the problem solving (Problem #1), describing what, who and when a determined action was executed. In addition, it is possible to see the quantitative and qualitative variables values for each group's member.

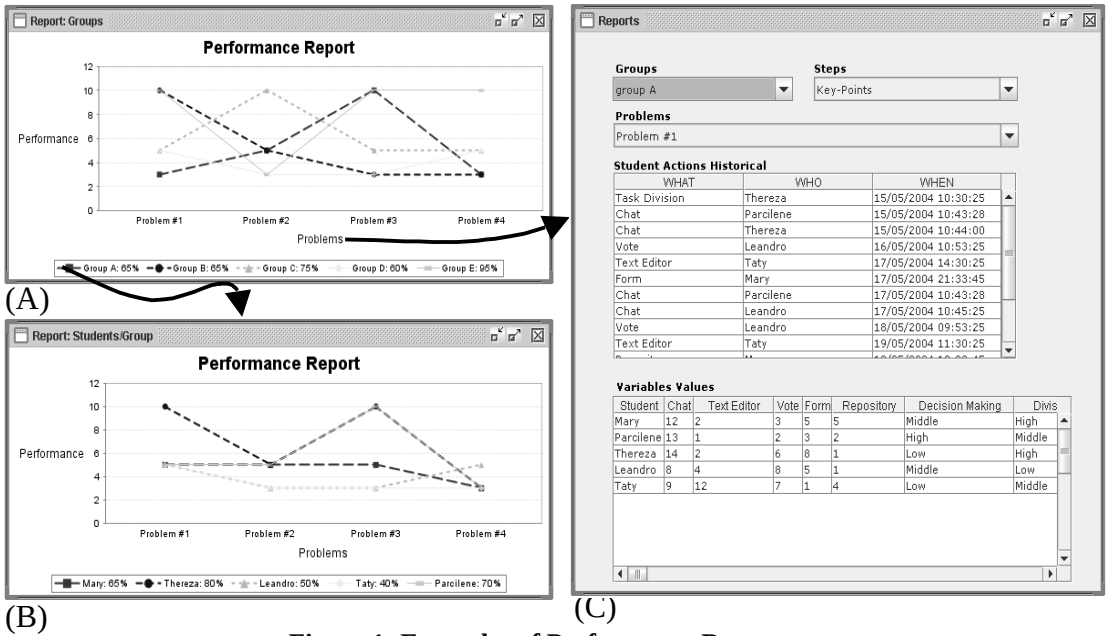

Figure 1: Examples of Performance Reports

The performance reports offer a multiple perspective view of the collaborative learning process. Group performance reports simplify the nature of interaction among existing groups. Student performance reports help to explain because a specific group had a below performance (value 2), for example. The report's representation form, graphic, facilitates also its general analysis. Moreover, there are recourses available to query quantitative and qualitative variables values for each student.

4 Conclusion

The manner to capture and store student actions in joint process is essential for designing of consistent collaborative models. The paper presented how mining techniques can help in the processing student interactions data and determining of the learning performance. The text mining technique can be seen as an alternative for understanding of conversations patterns among students, without that they need to define their contribution types via sentence openers avoiding a possible error. The categorization of contribution types by awareness agent has demonstrated a reasonable performance because the natural language has many ambiguity and still need of human interpretation. In our model, it is necessary to build a complete set of categories. For each existing category, many words and terms should be associated. The data mining technique used, semi-supervised learning, is very useful to compare the current state of interaction to ideal state because of the possibility for background knowledge incorporation.

According to realized experiments, the model has shown some relevant results for analyzing student performance. The results are not totally accurate because we do not have still a real data set that expresses information related to problem solving. Some refinements are being realized to improve the semantic analysis in messages and offer mechanisms for supporting inferences.

Topics for Discussion in the Workshop

Our group is working in the designing of computational models of collaborative learning interaction using mining techniques. With this perspective, we have some questions for discussion in the workshop:

- Are there any advantages for the definition of standard representation of collaborative models? Robustness? Extensibility? Information interchange? Interactivity? Is there already an initiative for standardization of the model?
- Recently, XML bas been proposed as the standard data representation for many applications. What are some advantages and disadvantages of using XML as a collaborative learning representation language? Only for possible access in the elements of the model?
- Which steps a system should be able to perform with interactions data before building models? Data cleaning? Selection of relevant examples? Treatment with missing values? Any methodology?
- From individual interaction data set is it possible to predict the future group performance? To possibility new inferences? Uncertainty?
- Why is it important that the user does not intervene in the categorization of the dialogue?
- What are the problems that have to be addressed when using TM technique for categorizing texts in the context of CSCL applications?

References

Bruce, R. (2001). Semi-supervised Learning Using Prior Probabilities and EM. Proceedings of the IJCAI Workshop on Text Learning. pp. 17-22.

Fayyad, U., Shapiro, G. P. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AAAIMIT Press, pp.37-54.

Katz, S., Aronis, J. and Creitz, C. (1999). Modelling Pedagogical Interactions with Machine Learning. In S.P. LaJoie & M. Vivet (Eds.), Proceedings of the Ninth International Conference on Artificial Intelligence in Education, LeMans, France, pp. 543-550.

Jermann, P., Soller, A., and Muehlenbrock, M. (2001). From Mirroring to Guiding: a Review of State of the Art Technology for Supporting Collaborative Learning. pp. 324-331

Martinez, A., Marcos, J.A., Garrachon, I., Fuente, P. and Dimitriadis, Y. (2002) Towards a Data Model for the Evaluation of Participatory Aspects of Collaborative Learning. CSCL 2002 Workshop: Designing Computational Models of Collaborative Learning Interaction, Boulder, Colorado.

McManus, M. and Aiken, R. (1995). Monitoring Computer-based Collaborative Problem Solving. Journal of Artificial Intelligence in Education 6, 4, pp. 307-336.

Quinlan J. R. (1996). Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research: pp.77-90.

Padilha, T. P. P. (2003). Um Ambiente de Aprendizado Colaborativo para Resolução de Problemas, Monografia de Qualificação de Doutorado, CPGCC, UFSC.

Pol, J. (2002). Identifying and Modeling Variables in Complex CSCL-situations - Case Study: The Use of Asynchronous Electronic Discussions. CSCL 2002 Workshop: Designing Computational Models of Collaborative Learning Interaction, Boulder, Colorado.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47.

Soller, A. and Lesgold, A. (2000). Modeling the Process of Collaborative Learning. International Workshop on New Technologies in Collaborative Learning, Awaji-Yumebutai, Japan.

Tan, A. H. (1999). Text Mining: The State of The Art and The Challenges. In Proceedings of the PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, pp. 65-70.

Vapnik, V. (1998). Statistical Learning Theory. New York: John Wiley&Sons.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning – ICML, Williamstown, pp. 577-584.

Williamson, J. (2002). Maximizing Entropy Efficiently. In the 19th Workshop on Machine Intelligence, Imperial College at Wye, Department of Philosophy, King's College London.