Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering

G. COBO, D. GARCÍA, E. SANTAMARÍA, J.A. MORÁN,
J. MELENCHÓN, C. MONZO
Universitat Oberta de Catalunya (UOC), Spain

Online discussion forums (or discussion boards) are one of the most common tools in web-based teaching-learning environments. Students' activity in discussion threads can be a relevant source of information that facilitates the monitoring tasks during the course by providing teachers with relevant indicators of their students' needs and lacks. In the present paper, the use of time series and an agglomerative hierarchical clustering algorithm is proposed with the aim of determining what different behavior patterns are adopted by students in online discussion forums. To this end, the actions carried out by students along the threads (e.g., writing and reading) are used to represent their activity in times series form. The use of an agglomerative hierarchical clustering algorithm is proposed in order to group similar students according to their activity profile. Some strategies on how to cut the obtained dendrograms are discussed and preliminary experimental results are presented.

Key Words and Phrases: Online discussion forums, data mining, clustering educational data, agglomerative hierarchical clustering, modeling students' activity

1. INTRODUCTION

Online discussion forums (or discussion boards) are emerging as one of the most common tools in web-based teaching-learning environments. On the one hand, an online discussion forum, being an asynchronous tool, allows students to maintain discussions related with their learning processes at any time. On the other hand, it also allows teachers to develop monitoring an even assessment tasks. In fact, a high level of interaction among students is desirable and increases the effectiveness of distance education courses (Fulford & Zhang, 1993). Then, all the activity carried out in the discussions threads appears to be a relevant source of information that can provide teachers with relevant indicators of their students' needs and lacks.

The purpose of the present work is to propose a strategy in order to model students' activity in online discussion forums. In such a context, clustering students seems to be a proper way to find similar learning behaviors (Vellido *et al.*, 2009). Thus, the proposed strategy is based on grouping students according to their activity in the online discussion threads, in order to identify what similar behavior profiles can be found in the virtual classroom. Due to the number of profiles is a priori unknown (in fact, it can depend on several factors: total number of students, kind of teaching-learning strategies promoted by the teacher, kind of subject, etc.) an agglomerative hierarchical clustering algorithm is used in order to group the students and find the behavior profiles.

This paper is structured as follows. Firstly, the working framework is introduced in Section 2. Next, the proposal of a novel strategy to model students' activity in online discussion forums is addressed in Section 3. Preliminary experimental results are shown in Section 4. Finally, conclusions and further work are presented in Section 5.

Authors' addresses: G. Cobo, D. García, E. Santamaría, J.A. Morán, J. Melenchón and C. Monzo, Informatics, Multimedia and Telecommunications Department, Universitat Oberta de Catalunya (UOC), Barcelona, Spain. E-mails: {gcobo, dgarciaso, esantamaria, jmoranm, jmelenchonm, cmonzo}@uoc.edu

2. WORKING FRAMEWORK

The application of data mining techniques in web-based teaching-learning environments is an incipient area which can solve many problems or support teachers in their decisions (Romero *et al*., 2006). Regarding student modeling field, different approaches are possible depending on the source of the analyzed data (logs, assessment and performance, tailored questionnaires, discussion threads, etc.), the kind of data mining techniques used in the modeling process (classification techniques, clustering algorithms, association rules, etc.) and the final application (predicting performance, identification of gifted students, implementing adaptive learning systems, etc.) (Romero *et al*., 2010).

This paper is focused on modeling students' activity in online discussion forums. In this regard, relevant contributions can be found in literature. For instance, an analysis of different styles of posting is made in (Sackville and Sherratt, 2006) from both quantitative and qualitative perspectives. For such purpose, four different types of messages are considered: statement, limited response, questioning response and dialogue postings. In a similar way, a system that identifies discussion threads that may have unanswered questions and need instruction attention is built in (Ravi and Kim, 2007), where messages are classified in four different categories: questions, answers, elaborations and corrections. In more qualitative terms, a taxonomy of participation in online discussions based in two variables –interpersonal interaction and interaction with content– is proposed in (Bento, 2005).

Moreover, in terms of modeling students' behaviors in online asynchronous environments, (Beaudoin, 2002) deals with identification of lurkers (in an online discussion board, a lurker is the one who reads but never writes). In online teaching-learning environments, this kind of behavior makes impossible a visible and active interaction with other students and teacher. In order to investigate lurking, (Nonnecke and Preece, 2001) carried out a study on lurking using in-depth semi-structured interviews with members of online groups. The analysis reveals that lurking is a strategic activity involving more than just reading posts. Reasons for lurking are categorized and a model is proposed to explain lurker behavior. Worker, lurker and shirker students are identified in (Taylor, 2002), after finding three significant participation patterns –proactive, peripheral and parsimonious participation groups, respectively– in accessing and contributing to a discussion board.

In this scenario, a novel data mining-based strategy to model students' activity in online discussion forums is proposed in the next section of this paper.

3. MODELING STUDENT'S ACTIVITY IN ONLINE DISCUSSION FORUMS

Regardless of the different types of posting one can consider (Sackville and Sherratt, 2006) (Ravi and Kim, 2007), two main actions are carried out in an online discussion forum: writing and reading. Students can adopt different attitudes in terms of these two main actions, thus defining different behavior profiles (active learners, lurkers, etc.). Modeling students' activity in online forums consists on finding groups of students with similar behaviors in terms of the actions they carry out all along the discussion threads.

Then, on the one hand, the first question is: what representation of students' activity can properly reflect their behavior profile? One common approach would be just counting the amount of actions carried out by students (Taylor, 2002). Our proposal is to represent students' activity in times series form. Thereby, both amounts and tendencies in students' activity can be identified (e.g. it's not the same reading thirty posts in one single day than along one week). A time series that represents a single student's activity would be:

$$\mathbf{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)} \cdots x_n^{(i)} \cdots x_N^{(i)}\right)$$

where $\mathbf{x}^{(i)}$ is the time series that represents the *i*-th student, $x_n^{(i)}$ is the value $\mathbf{x}^{(i)}$ adopts at temporal sample *n* and *N* is the total number of temporal samples the time series includes (*N* weeks, *N* days, *N* hours…). In a in a virtual classroom with *M* students, several strategies can be carried out in order to define the value of $x_n^{(i)}$, e.g.:

$$x_n^{(i)} = \frac{rd_n^{(i)}}{rd^M}, \quad x_n^{(i)} = \frac{wr_n^{(i)}}{wr^M}, \quad x_n^{(i)} = \begin{cases} 1 & \text{if } rd_n^{(i)} \neq 0 \\ 0 & \text{else} \end{cases}, \quad x_n^{(i)} = \begin{cases} 1 & \text{if } wr_n^{(i)} \neq 0 \\ 0 & \text{else} \end{cases}$$

where $rd_n^{(i)}$ and $wr_n^{(i)}$ are respectively the number of reading and writing actions carried out by the *i*-th student at sample *n* and rd^M and wr^M are respectively the total amount of reading actions carried out by all *M* students (the whole virtual classroom).

And, on the other hand, another question arises: how to group students with similar behaviors (i.e. similar time series)? The result will be a certain number of clusters which contain learners with similar behavior profiles, but the number of clusters (i.e. the number of profiles) is a priori unknown. Besides, it seems logical setting up the groups by linking first the most similar couple of students, then linking the next most similar couple and so on. Given these premises, our proposal is to use an agglomerative hierarchical clustering algorithm, a well-used algorithm in educational data mining (Vellido *et al*., 2009).

This kind of clustering algorithm presents significant advantages in this scenario. Firstly, the number of clusters is not required a priori, since the outcome of the algorithm is a hierarchical tree called dendrogram (Jain and Dubes, 1988) that shows how data joins along a tree structure of link points (nodes). Secondly, algorithm's behavior is completely deterministic, since requires no initialization. Thirdly, several labels for the data can be obtained by cutting the dendrogram according to different criteria. The most common one fits a certain distance threshold (the nodes under the threshold remain joined). Even so, criteria based on fitting inconsistency thresholds (the most consistent links remain joined now) can provide with more interesting results (Zahn, 1971). And fourthly, the dendrogram is an excellent exploratory data tool: its hierarchical structure gives interesting information on how students have been grouped and where isolated clusters can be found. Furthermore, it allows analyzing relations between students joined under the same node (this can be very useful in order to define behavior sub-profiles).

Hierarchical clustering algorithms need to fix two parameters in order to be properly configured. The first one is the distance function used to compare data. In this regard, Euclidean, Minkowski and Cosine are the most common distance functions when comparing time series (Gan *et al*., 2007). It has to be taken into consideration that certain distance functions may require of certain pre-processing of the time series (normalization, noise removing, etc.). And the second one is the linkage method used to measure distance

Table I. Five-step strategy to modeling students' activity in online discussion forums

Steps	Parameters
1. Defining the data	• Defining set of M students • Defining type(s) of messages (all type, questions, answers...) • Defining type(s) of actions (writing, reading...)
2. Constructing time series	• Defining type and number of samples (N weeks, N days, N hours...) • Defining the value of the times series data
3. Pre-processing time series	• Normalization, denoising, linear trend removing… • Time domain, frequency domain, feature extraction…
4. Agglomerative Hierarchical Clustering of time series	• Defining distance function: Euclidean, Minkowski, Cosine… • Defining linkage method: Single Link, Complete Link…
5. Identifying clusters (behavior profiles)	• Visual identification of clusters • By cutting dendrogram: distance threshold, inconsistency threshold…

between clusters. Single Link (nearest neighbor method) and Complete Link (farthest neighbor method) are the most usual methods (Gan *et al*., 2007).

4. PRELIMINARY EXPERIMENTAL RESULTS

In order to illustrate the proposed strategy, preliminary results are presented. The experiment involves a virtual classroom of 55 online students in an Electronic Circuits Theory subject and covers an entire semester (369 writing and 14142 reading actions throughout 119 days). Students' writing and reading actions are separately characterized by two different time series, without distinguishing among different types of messages:

$$\mathbf{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)} \cdots x_n^{(i)} \cdots x_{119}^{(i)}\right) \rightarrow x_n^{(i)} = \begin{cases} 1 & \text{if } wr_n^{(i)} \neq 0 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{y}^{(i)} = \left(y_1^{(i)}, y_2^{(i)} \cdots y_n^{(i)} \cdots y_{119}^{(i)}\right) \rightarrow y_n^{(i)} = \begin{cases} 1 & \text{if } rd_n^{(i)} \neq 0 \\ 0 & \text{else} \end{cases} \qquad \forall i \in [1,55]$$

Furthermore, Cosine distance (seeking for different activity paces) and Complete Link algorithm (emphasizing clusters compactness) have been used to cluster the times series data in time domain. The results of the experiment are shown in figure 1:

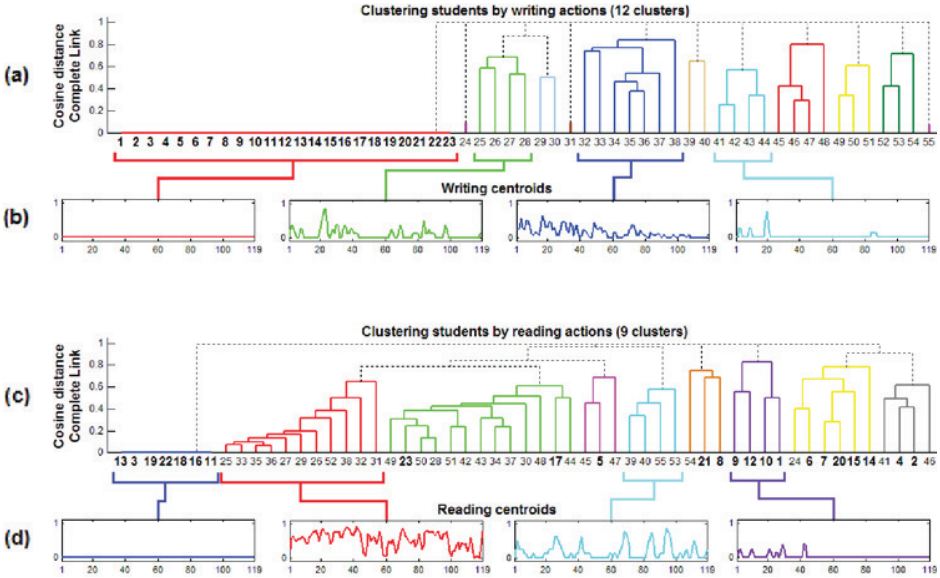

Fig. 1. (a): Writing clusters. (c): Reading clusters. (a) and (c): Bold-labeled students [1,23] are possible lurkers (no writing action). (b) and (d): Some representative profiles (centroids of some clusters) of both writing and reading throughout time are shown.

Some interesting remarks can be made from these results. Firstly, natural clusters (colored lines in Fig. 1 (a) and (c)) can be obtained by cutting dendrograms through suitable inconsistency thresholds (calculated according to (Zahn, 1971)), since these clusters are just hanging from the most inconsistent links (dotted lines in Fig. 1 (a) and (c)). In fact, the 9 clusters in Fig. 1 (c) cannot be obtained by cutting the dendrogram through any distance threshold, since there is no horizontal cut of the dendrogram that can provide with the same 9 clusters.

Secondly, resultant clusters centroids, both in terms of writing and reading (see Fig. 1 (b) and (d), respectively), indicate the most representative behavior profiles. Both total inactivity and regular activity profiles can be observed, as well as different profiles of more sporadic activity in different periods of time, for instance. This distinction based on different kinds of pace is possible because the Cosine distance tends to join profiles with coincident variations of activity throughout time.

And thirdly, students are grouped in different clusters depending on whether writing or reading actions are considered, which means that different behaviors can be associated to both kinds of activity. In fact, this strategy allows to easily distinguish between lurkers and inactive students: students #1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 17, 20, 21 and 23 perform a lurking behavior, since they don't write at all but they do have active reading profiles; whereas #3, 11, 13, 16, 18, 19 and 22 are truly inactive students, since they neither write nor read (see bold-labeled students in Fig. 1 (a) and (c)).

5. CONCLUSIONS AND FURTHER WORK

A novel strategy to model students' activity in online discussion forums has been proposed. This strategy is based on characterizing students' activity in time series form and grouping similar students through an agglomerative hierarchical clustering algorithm.

Time series allow representing different kinds of activity (e.g., writing and reading) and thus identify several activity profiles throughout time. Different groups of students can be obtained by means of agglomerative hierarchical clustering without knowing the number of clusters a priori. Furthermore, dendrograms are excellent exploratory data tools thanks to the rich information provided by their hierarchical tree structure.

Results obtained in a preliminary experiment point out that cutting dendrograms through a robust criterion based on links' inconsistency is the most proper way to obtain natural clusters. The obtained activity profiles also depend on the chosen distance function and on how the values of times series data are set. Furthermore, the proposed strategy seems to be really suitable to identify lurking behaviors.

In this sense, the work started in this paper will be continued. More experiments will be carried out in order to improve the modeling of students' activity. Moreover, student's actions will be labeled by adding qualitative information about threaded posts (questions, answers, corrections, etc.), in order to obtain more relevant behavior profiles.

REFERENCES

BEAUDOIN, M.F. 2002. Learning or lurking? Tracking the 'invisible' online student. *The Internet and Higher Education*, 2002, 5, 2, 147-155.

FULFORD, C.P. AND ZHANG, S. 1993. Perceptions of interaction: The critical predictor in distance education. *The American Journal of Distance Education*, 1993, 7(3), 8-21.

GAN, G., MA, C. AND WU, J. 2007. Data clustering. Theory, algorithms and applications. *ASIA-SIAM Series on Statistics and Applied Probability*, 2007.

JAIN, A., AND DUBES, R. 1988. Algorithms for Clustering Data. *Prentice-Hall*, 1988.

NONNECKE, B. AND PREECE, J. 2001 . Why lurkers lurk. *Americas Conf. on Information Systems*, 2001.

RAVI, S. AND KIM, J. 2007. Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers. In *Proceedings of the Conference on Artificial Intelligence in Education*. IOS Press, 357-364.

ROMERO, C., VENTURA, S. AND HERVÁS, C. 2006. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 2006, 33, 135–146.

ROMERO, C., VENTURA, S. PECHENIZKIY, M. AND BAKER, R. 2010. Handbook of educational data mining. *Data Mining and Knowledge Discovery Series,* Chapman & Hall / CRC Press, 2010.

SACKVILLE, A. AND SHERRATT, C. 2006. Styles of Discussion: Online Facilitation Factors. In *Proceedings of the Fifth International Conference - Networked Learning*, Lancaster University, 2006.

TAYLOR, J.C. 2002. Teaching and learning online: the workers, the lurkers and the shirkers. *Journal of Chinese Distance Education*, CCRTVU Press, Beijing, n. 9, 2002, 188, 31-37.

VELLIDO, A., CASTRO, F. AND NEBOT, A. 2010. Clustering educational data. In *Handbook of educational data mining*, CRC Press, 2010, 75-92.

ZAHN, C.T. 1971. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 1971, 20, 1.