Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums

Germán Cobo
IMT Department, Universitat Oberta de Catalunya (UOC)
Barcelona, Spain
+34 93 326 357
gcobo@uoc.edu

David García-Solórzano
IMT Department, Universitat Oberta de Catalunya (UOC)
Barcelona, Spain
+34 93 326 3686
dgarciaso@uoc.edu

Jose Antonio Morán
IMT Department, Universitat Oberta de Catalunya (UOC)
Barcelona, Spain
+34 93 326 3618
jmoranm@uoc.edu

Eugènia Santamaría
IMT Department, Universitat Oberta de Catalunya (UOC)
Barcelona, Spain
+34 93 326 3743
esantamaria@uoc.edu

Carlos Monzo
IMT Department, Universitat Oberta de Catalunya (UOC)
Barcelona, Spain
+34 93 326 3895
cmonzo@uoc.edu

Javier Melenchón
IMT Department, Universitat Oberta de Catalunya (UOC)
Barcelona, Spain
+34 93 326 3508
jmelenchonm@uoc.edu

ABSTRACT
Online discussion forums are a key element in virtual learning environments. The way learners participate in discussion boards can be a very useful source of indicators for teachers to facilitate their tasks. The use of a two-stage analysis strategy based on an agglomerative hierarchical clustering algorithm is proposed in this paper to identify different participation profiles adopted by learners in online discussion forums. Different parameters are used to characterize learners' activity (amount of posts, rhythm, depth of threads, crossed replies, etc). Participation profiles are identified and analyzed in terms of behavior and performance.

Categories and Subject Descriptors
J.1 [**Administrative Data Processing**] Education; K.3.1 [**Computer Uses in Education**] Distance learning;

General Terms
Algorithms, Measurement, Performance and Experimentation.

Keywords
Learning analytics, Educational data mining, Learner behavior modeling, Hierarchical clustering, Online discussion forums.

1. INTRODUCTION
Online discussion forums (or boards) are one of the most common tools in web-based teaching-learning environments. Online learner participation has been defined as a complex and intrinsic part of online learning [6]. In fact, a high level of interaction is desirable and increases the effectiveness of distance education courses [5]. Thus, discussion boards can be a relevant source of information in order to provide teachers with useful indicators of learners' activity and to facilitate their monitoring, guidance and feedback tasks.

The purpose of the present work is to present a two-stage analysis strategy in order to model and identify learners' participation profiles in online discussion forums. Since clustering learners has proved to be a proper way to find similar learning behaviors [11], learners with similar activity patterns are clustered together in the first stage and resultant clusters are combined in the second stage to identify participation profiles.

This paper is structured as follows. The working framework is introduced in Section 2; the clustering algorithm used in the experiments is proposed in Section 3; the data set is described in Section 4; the modeling strategy to identify participation profiles and the obtained results are shown in Section 5; and, finally, conclusions and future work are presented in Section 6.

2. WORKING FRAMEWORK
Relevant contributions can be found in literature on modeling learner behavior in online asynchronous environments. [1] deals with identification of lurkers (in a discussion board, a lurker is the one who reads but never writes). This kind of behavior makes impossible a visible and active interaction both with other learners and teacher in virtual environments. In order to investigate lurking, [8] 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 and a model to explain lurker behavior is proposed. Finally, three significant participation patterns in accessing and contributing to an online discussion board are defined in [10]: workers (proactive participants that are continuously involved in discussions), lurkers (peripheral participants that regularly access to the board and participate in the discussions in read-only mode) and shirkers (parsimonious participants that barely access to the board).

LAK'12, 29 April – 2 May 2012, Vancouver, BC, Canada.

A different approach is provided by social network analysis techniques, in order to show interactions between learners in online discussion threads (learner network) and evaluate their participation [9]. Moreover, a great diversity of indicators (depth of threads, rhythm, reciprocal readings, cross replies, etc.) is used from this approach in order to define effective interaction models capable of giving an immediate picture of the effectiveness level of a collaborative group [2].

Finally, interesting contributions on participation profiles in discussion boards of general topics –not strictly educational– can be found as well. Several online forums of different topics are classified in [3] regarding their predominant user roles rather than their topics. Eight different user roles (popular initiators, popular participants, joining conversationalists, supporters, taciturns, grunts, elitists and ignored) are identified through an analysis method based on PCA (the most dominant feature in the largest component is selected in order to define three bands of users and discard the lowest and middle ones –marginal participation profiles–) and agglomerative hierarchical clustering (the optimal number of clusters is selected after an inspection of the solutions provided by different validation techniques).

3. CLUSTERING ALGORITHM

The modeling strategy in the present work is based on identifying participation profiles from the different activity patterns conducted by learners in online discussion forums. In order to group learners with similar activity patterns together, a clustering algorithm is used [11]. Due to the number of relevant patterns (i.e., the number of relevant clusters) is a priori unknown, we use an agglomerative hierarchical clustering algorithm [3].

The outcome of an agglomerative hierarchical clustering algorithm is not a data partition, but a dendrogram-type graph [7]. A dendrogram is a hierarchical tree structure formed by links that join couples of clusters together in a new cluster (i.e., each link defines a possible cluster of data) from the beginning (singleton clusters –i.e., one cluster per learner–) to the end (a unique cluster including the whole set of data –i.e., all learners grouped together–) of the tree. The heights of the links correspond to the distance between the couple of clusters joined as a new cluster under the link. The similarity measure between clusters depends on the linkage function defined in the algorithm: the Single Link (nearest neighbor) and Complete Link (furthest neighbor) are the most popular linkage functions [7].

A dendrogram is a useful tool for both visually exploring similarities between data (data exploratory analysis) and obtaining data partitions (clusters of data). Classical strategies to get a data partition consist in cutting the dendrogram at any defined threshold height and dismiss the links above the cut [7]. More interesting is to evaluate links in terms of their inconsistency instead of their height and define a threshold inconsistency in order to dismiss the most inconsistent links [12]. Finally, more versatile strategies try to isolate clusters separately as the dendrogram grows, for the sake of flexibility and to be able to detect both sparse and dense clusters [4].

The agglomerative hierarchical clustering algorithm used in this paper combines the strategy of isolating clusters separately (instead of getting a final data partition in one go by a single cut in the dendrogram) with a modified version of the inconsistency criterion defined in [12]. Our algorithm builds the whole dendrogram and isolates its best cluster in terms of a consistency criterion (the best cluster is the one defined by the most consistent link). Once the best cluster is isolated, this process is iterated until there is no remaining data to be isolated.

Thus, taking the *gap* concept (height increment between consecutive links) proposed in [4], we define:

$$c_i = \left[z_i - \max\left(u_i\right)\right] \cdot k\left(n_i\right)$$

as the consistency of the i-th link in the dendrogram, being:

$$z_i = \frac{gap_i - \mu_i}{\sigma_i}$$

$$k\left(n_i\right) = 1 - e^{-\alpha \cdot n_i} \ni \alpha = \frac{-100}{\beta \cdot N_{TOT}} \cdot \ln\left(1-\gamma\right), \forall \gamma \in \left]0,1\right[$$

where gap_i is the gap above i-th link, μ_i and σ_i are the mean and the standard deviation of the population formed by gap_i and the gaps above all the links nested under the i-th link (z_i is the standard score of gap_i), u_i is the set of standard scores of the gaps above all the links nested under the i-th link, N_{TOT} is the total amount of elements (i.e., learners) in the data set, n_i is the amount of elements within the cluster defined by the i-th link and $k(n_i)$ is an exponential correction applied to avoid isolating too small size clusters ($k(n_i)$ is less than γ when n_i is less than the β % of N_{TOT}).

4. DATA SET

The experiments conducted in this paper analyze the activity carried out by learners within the online discussion forums of three different subjects in a virtual Telecommunications Degree (Electronic Circuits, Linear Systems Theory and Mathematics) and throughout three complete semesters (from February 2009 to July 2010). All the courses took place in an asynchronous web-based teaching-learning environment and the participation of learners in discussion boards was not mandatory, but strongly recommended. Thus, the whole dataset involves a total amount of 672 learners (N_{TOT}) distributed in eighteen different virtual classrooms and a total amount of 3842 posts. Total withdrawal and passing rates are 36.31% and 52.23%, respectively.

5. MODELING ACTIVITY AND FINDING PARTICIPATION PROFILES

The analysis strategy conducted in the present work consists of two main stages. In the first stage, learners' activity in online discussion forums is characterized in two different domains (writing and reading) and learners with similar activity patterns are grouped together in each domain separately.

The activity carried out by learners is differently characterized in each domain (different parameters are used depending on the domain). In writing domain, each learner is characterized according the following four parameters (all of them are ratios over learner's specific virtual classroom and semester): ratio of threads –weighted by their respective depths– initiated by learner over total amount of threads –weighted, as well– (***depth***), ratio of reply posts written by learner over total amount of reply posts (***re***$_{posts}$), ratio of learners replied –at least, once– by learner over total amount of learners (***re***$_{cross}$) and ratio of days when learner

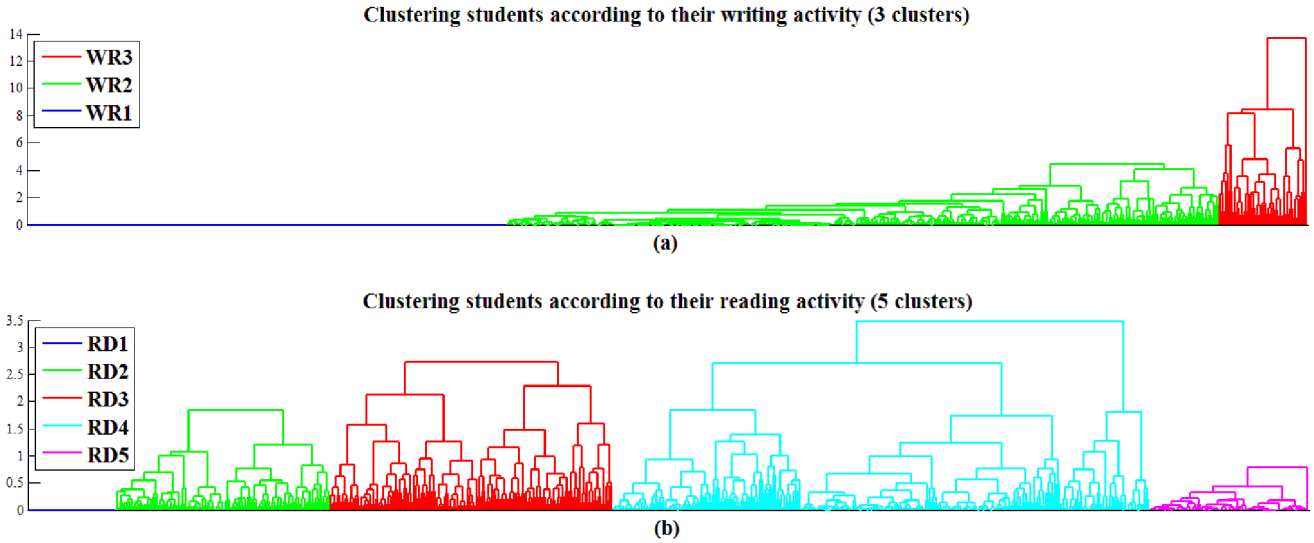

Figure 1. This figure shows the resulting dendrograms from clustering learners in terms of (a) writing and (b) reading. Each dendrogram legend indicates the assigned label to each cluster, which can be identified by its color. According to the criterion defined by the algorithm, resultant clusters are nested under the most consistent links in the dendrogram. Each cluster top height corresponds to the distance between the furthest learners within the cluster (Complete Link).

Table 1. This table shows how the participation profiles are identified. Combining the resultant clusters from the first stage of analysis, the final set of clusters is obtained (i.e., learners belonging to WRi and RDj clusters belong now to the new WRi–RDj cluster). In the table, final clusters are represented in rows, first column show the final clusters labels, *N* column (% over N_{TOT}) indicates the amount of learners per cluster, Withdrawal and Passing columns (% over *N*) indicate withdrawal and passing rates at the end of semester, Average Representatives (Centroids) columns show each final cluster's average representative (in terms of the eight different parameters used to characterize learners' activity), and the last two columns describe the participation profiles represented by each final cluster (by identifying the different participation profiles proposed in [10]).

Final Clusters	*N*	Withdrawal	Passing	Average Representatives (Centroids)								Participation Profiles	
				depth	re_{posts}	re_{cross}	wr_{rhythm}	rd_{posts}	$rd_{threads}$	rd_{cross}	rd_{rhythm}		
WR3–RD5	4.17	0	96.43	0.18	0.21	0.27	0.82	0.99	1	1	0.94	Workers	High-level Workers
WR3–RD4	2.68	5.56	83.33	0.25	0.19	0.18	0.79	0.9	0.93	0.9	0.68		
WR2–RD5	7.29	4.08	89.8	0.03	0.04	0.07	0.22	0.99	1	1	0.95		Mid-level Workers
WR2–RD4	29.76	19.5	65	0.03	0.03	0.05	0.19	0.8	0.84	0.9	0.56		
WR2–RD3	12.95	44.83	41.38	0.01	0.01	0.02	0.12	0.36	0.45	0.66	0.25		Low-level Workers
WR2–RD2	5.51	45.95	37.84	0.01	0	0.01	0.07	0.1	0.17	0.31	0.09		
WR1–RD5	0.89	16.67	83.33	0	0	0	0	0.98	1	0.96	0.94	Lurkers	High-level Lurkers
WR1–RD4	9.52	26.56	59.38	0	0	0	0	0.82	0.86	0.89	0.49		
WR1–RD3	9.23	64.52	27.42	0	0	0	0	0.31	0.36	0.66	0.16		Low-level Lurkers
WR1–RD2	11.16	60	32	0	0	0	0	0.07	0.12	0.26	0.06		
WR1–RD1	6.7	93.33	2.22	0	0	0	0	0	0	0	0	Shirkers	
WR2–RD1	0.15	100	0	0	0	0	0.05	0	0	0	0		
WR3–RD3	0	–	–	–	–	–	–	–	–	–	–	–	
WR3–RD2	0	–	–	–	–	–	–	–	–	–	–		
WR3–RD1	0	–	–	–	–	–	–	–	–	–	–		

writes at least one post over total amount of days (wr_{rhythm}). Other four different parameters are used in the reading domain (self-readings are excluded): ratio of posts read by learner over total amount of posts (rd_{posts}), ratio of threads where learner reads at least one post over total amount of threads ($rd_{threads}$), ratio of learners read –at least, once– by learner over total amount of learners (rd_{cross}) and ratio of days when learner read at least one post over total amount of days (rd_{rhythm}).

Learners are separately clustered in both domains by using the agglomerative hierarchical clustering algorithm described in Section 3 with the following configuration: Normalized Euclidean Distance, Complete Link, $\beta = 10$ and $\gamma = 0.9$. Results obtained in both domains are shown in Figure 1.

Finally, the second stage of the analysis strategy consists in grouping together those learners belonging to the same clusters in both writing and reading domains. Thus, the final set of clusters that completely defines the different activity patterns and allows to identify the participation profiles of learners in online discussion forums is obtained (see Table 1). Participation profiles are mapped to final clusters by observing and comparing the values of the parameters that characterize the learners' activity patterns in each cluster. Final clusters' centroids allow to confirm the suitability of this mapping and to describe and characterize the participation profiles in more detail.

Some interesting remarks can be made from the obtained results. Regarding the agglomerative hierarchical clustering algorithm performance, it allows to find clusters of different size and density in both different domains (e.g., in Figure 1 (a), WR2 cluster is larger and denser than the smaller and sparser WR3).

Participation profiles like the ones describe in [10] can be easily identified by observing final clusters' centroids (see Table 1): shirkers (inactive learners) are grouped within WR1-RD1 and WR2-RD1 clusters (centroids with no kind, or negligible, activity at all); lurkers (only readers), within WR1-RD2/…/-RD5 clusters (centroids with no kind of reading activity and different patterns of writing activity); and workers (active learners), within WR2-RD2/…/-RD5 and WR3-RD4/-RD5 clusters (different patterns of both writing and reading activity). Furthermore, specific sub-profiles for lurkers (low- and mid-level lurkers) and workers (low-, mid- and high-level workers) have been defined depending on differences between centroids' values of reading and writing parameters, respectively.

Empty possible combinations (WR3-RD1/…/-RD3) are also useful to deduce some –pretty logical– conclusions: writing involves reading, but not the other way around (reading does not necessarily involve writing –lurking behavior–). Besides, differences between centroids' values can be useful to identify other kinds of participation profiles as well (e.g., the different user roles proposed in [3]: popular initiators, popular participants, joining conversationalists, supporters, taciturns, elitists, grunts and ignored).

Finally, some interesting conclusions regarding on performance differences between profiles can be pointed out: the withdrawal rates of shirkers and low-level lurkers are the top highest and the passing rates of high-level lurkers are comparable with the ones of high- and mid-level workers' (which are logically the top highest).

6. CONCLUSIONS AND FUTURE WORK

In this paper, a two-stage strategy in order to model learner participation profiles in online discussion forums is proposed. The presented agglomerative hierarchical clustering algorithm successfully isolates the more relevant activity patterns in different domains (writing and reading). The obtained final clusters actually group learners with similar activity patterns and allow to satisfactorily identify different participation profiles in online discussion forums. In terms of future work, the number of domains in first analysis stage will be increased (rhythm domain, neighboring domain, etc.) and the impact of this increasing on both the presented clustering algorithm suitability and the identification of participation profiles accuracy will be checked.

7. REFERENCES

[1] Beaudoin, M.F. 2002. Learning or lurking? Tracking the 'invisible' online student. *The Internet and Higher Education*. 5, 2 (Jul. 2002), 147-155.

[2] Calvani, A., Fini, A., Molino, M. and Ranieri, M. 2009. Visualizing and monitoring effective interactions in online collaborative groups. *British Journal of Educational Technology*. 41, 2 (Mar. 2010), 213-226.

[3] Chan, J, Hayes, C. and Daly, E. 2010. Decomposing discussion forums and boards using user roles. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line* (Raleigh, NC, USA, April 23 - 24, 2010).

[4] Fred, A.L.N. and Leitão, J.M.N. 2003. A new cluster isolation criterion based on dissimilarity increments. *IEEE Transactions of Pattern Analysis and Machine Intelligence*. 28, 8 (Aug. 2010), 944-958.

[5] Fulford, C.P. and Zhang, S. 1993. Perceptions of interaction: The critical predictor in distance education. *The American Journal of Distance Education*. 7, 3 (1993), 8-21.

[6] Hrastinski, S. 2008. What is online learner participation? A literature review. *Computers & Education*. 51, 4 (Dec. 2008), 1755-1765.

[7] Jain, A. and Dubes, R. 1988. Algorithms for Clustering Data. *Prentice Hall*. 1988.

[8] Nonnecke, B. and Preece, J. 2001. Why lurkers lurk. In *Proceedings of Americas Conf. on Information Systems* (Boston, MAS, USA, August 03 - 05, 2001).

[9] Rabbany, R., Takaffoli, M. and Zaïane, O.R. 2011. Analyzing participation of students in online courses using social network analysis techniques. In *Proceedings of 4th International Conference on Educational Data Mining* (Eindhoven, The Netherlands, July 6 – 8, 2011).

[10] Taylor, J.C. 2002. Teaching and learning online: the workers, the lurkers and the shirkers. *Journal of Chinese Distance Education*. 9 (2002), 31-37.

[11] Vellido, A., Castro, F. and Nebot, A. 2010. Clustering educational data. In *Handbook of educational data mining*, CRC Press, 2010, 75-92.

[12] Zahn, C.T. 1971. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 20, 1 (Jan. 1971), 68-86.