Modelling and identifying collaborative situations in a collocated multi-display groupware setting

Roberto Martinez[1], James R. Wallace[2], Judy Kay[1], Kalina Yacef[1]

[1] School of Information Technologies, University of Sydney, NSW 2006, Australia
{roberto, judy, kalina}@it.usyd.edu.au
[2] Department of Systems Design Engineering, University of Waterloo, ON, Canada
jrwallac@uwaterloo.ca

Abstract. Detecting the presence or absence of collaboration during group work is important for providing help and feedback during sessions. We propose an approach which automatically distinguishes between the times when a co-located group of learners, using a problem solving computer-based environment, is engaged in collaborative, non-collaborative or somewhat collaborative behaviour. We exploit the available data, audio and application log traces, to automatically infer useful aspects of the group collaboration and propose a set of features to code them. We then use a set of classifiers and evaluate whether their results accurately match the observations made on video-recordings. Results show up to 69.4% accuracy (depending on the classifier) and that the error rate for extreme misclassification (e.g. when a collaborative episode is classified as non-collaborative, or vice-versa) is less than 7.6%. We argue that this technique can be used to show the teacher and the learners an overview of the extent of their collaboration so they can become aware of it.

Keywords: Data Mining, Group Modelling, Collaborative Learning

1 Introduction and related work

There are significant learning benefits of collaboration when students work in small groups [1]. However, in practical classroom settings, it is challenging for the teacher to be aware of the level of collaboration in each small group within their class. Emerging uses of technology offer the possibility of automatically capturing data that then can be used to detect the level of collaboration of a group. There are several ways in which such models of collaboration might be used, including mirroring information of the group to the learners and their teachers or improving the provision of adequate support in computer-supported collaborative learning systems [2]. In the latter case, these environments are sometimes designed to encourage learners to collaborate or present a structured task that forces collaboration and participation awareness. However, a general issue in applying these strategies is that different types of supportive actions can have different effects on the learning processes [3]. Specifically in collaborative learning environments, it has been shown that help is more effective if delivered just when it is needed [4]. Otherwise, well functioning groups may be distracted by unnecessary system interventions. Meanwhile, groups who experience problems and do not collaborate may benefit from such interventions.

The goal of our work is to explore ways to exploit readily available data to determine the level and nature of collaboration. This paper proposes an approach to infer whether group members are involved in collaborative behaviour or not. We make use of two forms of data. One is the presence of speech, based on an audio feed from each learner, without analysis of what is said. We call this a *simple audio trace*. The second source of data comes from the application log traces. From these two sources, we automatically infer key aspects of collaboration and propose a set of features to encode them. These features are then evaluated with a range of classifiers.

A given situation can be considered as "collaborative" in a learning context, if there are particular forms of interaction among the group members. For example, learning mechanisms such as explanation, negotiation, disagreement or elicitation [1]. However, even if the conditions under which these special interactions are present, there is no guarantee that learning will occur. We hypothesise that *it is possible to automatically infer whether a group of learners is engaged in a collaborative situation, from the application and audio traces of interaction with a reasonable level of accuracy.*

A number of research projects have analysed the interactions between learners to improve instructional support for collaboration using machine learning and user modelling techniques. In [5] the authors presented a fuzzy model for predicting forms of collaboration regarding the quality of the final group solution. Sequence pattern mining and clustering techniques were used to extract patterns and gain insights into the key factors that distinguish successful teams [6]. Additionally, supervised and unsupervised learning techniques have been used for grouping students according to their collaboration, assigning a value to each student to support comparison of students' behaviour [7]. The work in this paper breaks new ground as it focuses on mining patterns from simple audio and logs of interaction to match qualitative observations of the presence or absence of collaboration in a collocated group.

In the next section, we present related work. In Section 3, we introduce the collaborative learning context of our work, describing data collection and preparation. Section 4 presents our feature model, followed by the results of a number of learning approaches and we conclude with reflections and further work.

3 Context of the study and data exploration

The purpose of this research study was to explore whether it is possible to infer with a reasonable level of accuracy the level of collaboration within small groups of learners. We first present the environment in which our data was collected.

Data Collection. A previous study explored the impact of alternative shared displays on group processes [8]. Data was collected from 13 groups, each with 3 students, for a total of 39 students (Figure 1, right). The participants were students predominantly enrolled in university Maths, Science or Engineering courses and aged 18-27 years. Groups were asked to perform the Job Shop Scheduling (JSS) task, an optimisation problem specifically designed for evaluating interactions within groups of learners. Participants were asked to optimise the scheduling of six *jobs*, each composed of six ordered operations. These operations require the use of six resources that can only be in use by one operation at a time. Participants modify the interface by dragging *resource pieces* into position with the shared goal of scheduling the completion of all six jobs in a minimal amount of time (see Figure 1, left).

In addition to a large, shared display projected on a nearby wall, participants were provided with laptops and external mice through which they could perform individual actions. The interface visible on the personal laptops provided a personally tailored view of the workspace, where the resources that the owner could interact with were presented as more salient than the others. The large, shared display provided an overview of the group's task progress.

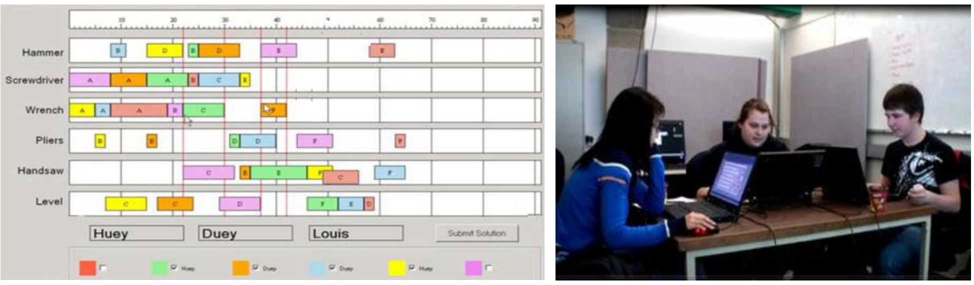

Fig. 1. Left: Application screenshot. Right: Group of students solving a problem.

Each group was required to develop solutions for the JSS task 2 or 3 times. Data from 29 trials were collected and coded. Groups spent 17 minutes per trial on average and executed between 100 and 600 *physical actions* per solution, for a total of 9,800 recorded mouse click or drag operations within the JSS software. In addition to the application logs, we also transcribed verbal utterances for each trial's video recording. Each complete unit of speech in spoken language produced by a learner was considered as a verbal participation. In general, most of groups' speech was on-task. These transcripts included a total of 4,836 *verbal participations* which, combined with the physical action data, formed a dataset of more than 14,636 physical and verbal interactions (Table 1 illustrates example logs of this dataset).

Table 1. Samples from the JSS combined dataset.

Verbal Participation Log				Physical Action Log			
User	Start	End	Log	User	Start	End	Log
C	15:18	15:19	I'll take care of the a's	A	01:57	01:59	- Move resource A-
C	15:21	15:22	you do the c and the d's	C	01:58	01:59	- Move resource D -
A	15:22	15:23	Yea	B	02:02	02:04	- Move resource B-

Data exploration. Before any data mining technique was performed, the data was examined to see whether any simple statistics could distinguish interesting differences between groups. Firstly, we calculated the total number of utterances, clicks and talking time for each group. Figure 2 shows the participation sequence diagrams of three sample groups. The top of each diagram shows the verbal participation and the lower parts represent the physical actions. The horizontal lines and rectangles represent actions or sets of actions (rectangles) performed by each author. The directed arrows indicate the relative sequence of the actions. From these diagrams, we observe that Group "K" was generally collaborative but participants A and C were more active. Group "L" did not have much verbal interaction, and from the diagram of physical actions, we observe they did not do much neither. Group "M" presents asymmetrical group activity: Student C has just three verbal actions, far less than the others in the group, but he performed most of the physical actions. These diagrams illustrate significant differences that exist between groups. These observations were confirmed by analysing the video recordings of the sessions.

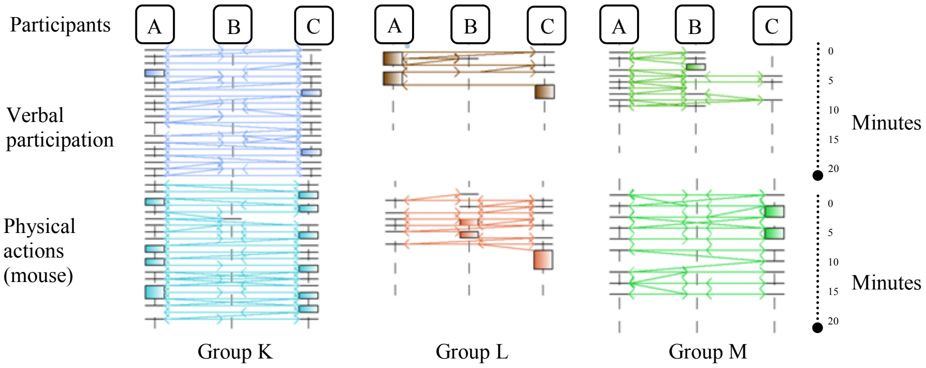

Fig. 2. Representation of the verbal and physical participation of three groups. A participative group (left), a non-communicative group (centre) and an asymmetric group (right). Diagrams created using the Process Mining Framework [9].

4 Learning collaborative behaviour

We now describe the rest of our approach, which, after collecting the logs of activity consists of annotating the data, constructing a set of features to learn these labels and applying different classifiers. Results are presented in the next section.

Data annotation. Dillenbourg [10] describes a situation as collaborative when participants are at the same level, can perform the same actions, have a common goal and work together. Building on these criteria, qualitative observations were made to assess whether each group was collaborating. Videos of each group's sessions were observed. Groups' activity was coded every 30 seconds based on the perception of collaboration for that block of time (as if a teacher was observing the group).

Each block of activity was coded as matching one of three possible values, the highest being a *collaborative moment* (C), based on Dillenbourg's definition of collaboration [10] described above. If all participants participated to some extent or they were aware of their peers' actions, then, that 30 seconds block of activity was tagged as "collaborative". A moment was tagged as *somewhat collaborative* (SC) if one or two members were unaware of their peer's actions, or if the group failed to communicate but they still tried to collaborate at some level. The last possible value, *non-collaborative moment* (NC), was assigned if the group split the task, working separately, or if just one participant did all the work. A label was assigned to each 30 seconds block of activity for each group. Most of the observations were carried out by a single observer. Two different raters, including a domain expert, tagged a sample of

15% of the sessions. Inter-rater reliability was reasonably acceptable – Cohen's k = 0.69. All groups had the same time to solve the problem (20 minutes) but they were free to decide when to stop. Figure 3 (left) depicts examples of the coding of some sessions. A row with many blue blocks (C), some in orange (SC) and few in light yellow (NC) corresponds to a collaborative sessions.

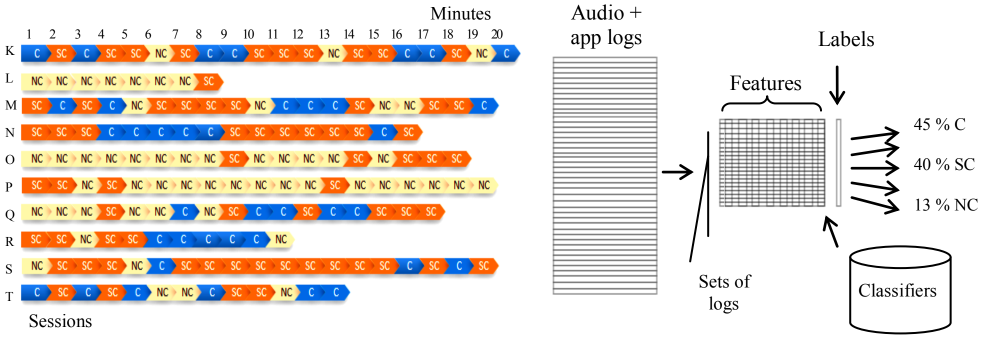
Minutes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
K C SC C SC SC NC SC C C SC SC SC NC SC SC C C SC NC C
L NC NC NC NC NC NC NC SC
M SC C SC C NC SC SC SC SC NC C C C SC NC NC SC SC C
N SC SC SC C C C C C SC SC SC SC SC SC C SC
O NC NC NC NC NC NC NC NC SC NC NC NC NC SC NC SC SC SC
P SC SC NC SC NC NC NC NC NC NC NC NC SC NC NC NC NC NC NC
Q NC NC NC SC NC NC C NC SC C C SC C C SC SC SC
R SC SC NC SC SC C C C C C NC
S NC SC SC SC NC C SC SC SC SC SC SC SC SC SC C SC C SC
T C SC C SC C NC NC C SC SC NC C C
Sessions
Audio + app logs
Sets of logs
Features
Labels
45 % C
40 % SC
13 % NC
Classifiers

Fig. 3 Left: Dot plot representations of the coding of some analysed sessions. Each 30 seconds of group work can be tagged as "collaborative" (blue), somewhat collaborative (orange) or Non-collaborative (yellow). Diagram created using the Process Mining Framework [9]. Right: The architecture of the collaborative model.

Then, the audio and application log lines were grouped forming sets of log lines (Figure 3, right). The grouping was done using three different block sizes: 30, 60 and 90 seconds. We chose these time frame sizes based on the observations made on the videos of the sessions. In a period of 30 seconds, we can observe complete dialogues related to a solution issue so we chose it as our minimal granularity. However, the conversations can last more than 30 seconds, so we also investigated the use of longer time-frames (60 and 90 seconds). For these, the label was obtained by implementing 60 and 90 second sliding windows with steps of 30 seconds and joining the underlying labelled blocks using the following rules when the labels were not uniform across the blocks: For 60 seconds: C+SC=C, SC+NC=NC,C+NC=SC. For 90 seconds: SC+SC+C= C, C+SC+NC= SC, NC+NC+*= NC, C+C+*=C, etc. Using this process, we obtained three datasets of similar size (700 samples in average).

Feature selection. Weinberger and Fischer [11] defined that two dimensions of the collaborative learning work that can be measured quantitatively are the amount and the heterogeneity of participation. Drawing on this, a number of features were calculated for each block. We propose a feature model that includes: quantity of physical and verbal participation (features 1, 2 and 3 in Table 2), number of active participants (feature 4) and the degree of dispersion of the participation among (features 5, 6 and 7) them. In this way, we obtained three different datasets in which each instance corresponds to one block of log lines grouped in 30, 60 or 90 seconds blocks. Speech recognition was *not* used in the analysis. If there were reliable recognition of speech, this might be fed into our approach. We used the Gini coefficient as an indicator of dispersion of participation as it has been successfully used to measure equity of participation in face-to-face collaborative settings [12]. For this coefficient, a value of zero means total equality and a value of one indicates maximal inequality.

Table 2. Diagnosis features and six examples of 30 seconds blocks of collaborative (C1, C2), somewhat collaborative (SC1, SC2) and non-collaborative (NC1, NC2) activity.

Feature	Metric	C1	C2	SC1	SC2	NC1	NC2
1-Physical participation	Scalar	7	12	15	15	10	15
2- Number of utterances	Scalar	28	9	4	5	0	4
3-Talking time	Seconds	19.4	17	7.5	8	0	5
4-Number of talking participants	0, 1, 2 or 3	3	3	2	3	0	1
5-Talking time dispersion	Gini coeff.	.510	.284	.747	.60	1	1
6-Verbal participation dispersion	Gini coeff.	.357	.143	.75	.614	1	1
7-Physical participation dispersion	Gini coeff.	.875	.583	.533	.40	.2	.8

5 Evaluation

We created classification models based on the three datasets described above. We used the Best-First tree, C4.5 decision tree, Bayes-Net and naïve Bayes algorithms. Similar techniques have been successfully applied in learning contexts for detecting behaviour patterns [13]. The models were evaluated using two methodologies: 10x 10-fold Cross Validation (CV) and Leave-one-group-out CV. The 10 runs of 10-fold CV were performed on each of the 3 datasets for each algorithm. This is equivalent to breaking the data into 10 sets of same size, training on 9 of them and testing on the 10th, repeating this 10 times (folds) and repeating the whole process also 10 times. We used a standard baseline for comparing the performance of the classifiers. The *baseline classifier* simply takes account of the distribution of the frequency of the three possible label values.

We obtained results that are significantly higher than the standard baseline (Table 3). In general, even when the accuracy of the models is above our baseline we obtained sub-optimal performance with all the algorithms to predict *somewhat collaborative situations* (SC row). The training dataset formed by blocks of 30 seconds produced some of the higher performance rates across the datasets (68% for naïve Bayes, 66% for Best-First tree and Bayes-Net of F-score), and it is more balanced in the prediction of the 3 possible values. For the second dataset, we got lower rates of performance compared with the others. The third dataset produced also high rates of correct predictions (F-score above .68 for the decision trees). However, an additional metric for gaining insights on the accuracy of the models was calculated. We call it *extreme misclassifications* (EX). This measures the proportion of incorrect classifications in which the *non-collaborative* blocks were misclassified as *collaborative* and vice versa. In educational terms, a *collaborative* block misclassified as *somewhat collaborative* is still giving information about the group activity. The proportion of extreme misclassifications for the 30 seconds dataset, for all the classifiers, stayed below 7.6%; therefore the results of these models are highly acceptable. For the 90 seconds dataset, even when the accuracy levels are comparable to the first dataset, it does not perform well with the extreme misclassifications (highlighted row). The row OP (Optimistic accuracy) shows what the accuracy levels would be if only extreme misclassifications are counted as errors. Whilst this is not ideal, it shows that the classifier model is very reliable. The algorithms which produce simpler models are the decision trees. Based on these we found that the features that define the most of the classification are the number of utterances produced (at least 10 for a collaborative situation, 30 sec. dataset), low rates of verbal participation dispersion (Gini coefficient less than .40) and the absence of long periods of silence.

Table 3. Results of the 10-fold cross validation. F1=Balanced F-score, C= F-measure of the algorithm in classifying "collaborative" SC=somewhat collaborative, or NC= non-collaborative situations. EX= extreme misclassifications accuracy, OP= optimistic accuracy. BL=baseline, BNet= Bayesian Network, NB= naïve Bayes, BFT= Best-first tree, C4.5= C4.5 tree.

	Log sets of 30 seconds					Log sets of 60 seconds					Log sets of 90 seconds				
	BL	**BNet**	**NB**	**C4.5**	**BFT**	BL	BNet	NB	C4.5	BFT	BL	BNet	NB	C4.5	BFT
F1	.340	**.656**	**.683**	**.625**	**.659**	.360	.659	.668	.638	.646	.360	.666	.624	.686	.682
C	.310	**.701**	**.709**	**.652**	**.659**	.460	.827	.860	.771	.821	.340	.771	.826	.656	.587
SC	.330	**.558**	**.576**	**.630**	**.572**	.220	.178	.223	.369	.248	.430	.496	.412	.695	.724
NC	.370	**.725**	**.787**	**.739**	**.720**	.330	.771	.722	.650	.691	.250	.808	.707	.713	.737
AC		.654	**.687**	.628	**.657**		.666	.716	.648	.695		.759	.719	.746	.732
EX	.280	**.045**	**.076**	**.072**	**.057**	.340	.452	.429	.472	.510	.270	.520	.538	.640	.646
OP	.820	**.984**	**.976**	**.973**	**.981**	.790	.846	.855	.829	.819	.830	.826	.798	.799	.794

Table 4 summarises the results of the Leave-one group out CV. This analysis shows similar accuracy levels for each algorithm compared to the 10-fold CV but it also generates additional information regarding the performance of the models for each group. We noted above that the classifier algorithms produce more equilibrated results for the classification (C/SC/NC) grounding on the 30 seconds blocks dataset. However, using a Leave one out approach on this dataset, we can notice how the accuracy falls or rises depending on the group that is being tested each run. The Bayesian algorithms (at least the Bayesian network) have less oscillation in the classifications in the 30 seconds dataset (std = 9.9%) compared with the decision trees algorithms (std = 14% and 13.5%). We analysed the correlation between the proportion of *collaborative moments* and how well each model performs (accuracy). We expect not to have high correlation between the accuracy and the proportion of collaborative moments. In table 4 we can see that the negative correlation increases for the SC blocks (below -.550 for Bayes-Net and naïve Bayes in all datasets). In other words, both Bayesian algorithms are good when groups clearly behave as very collaborative or non-collaborative and decrease their power for somewhat collaborative groups. In this same respect the trees performs "better" (corr. of -.34 and -.38 for the 30 sec. dataset) but their power of prediction oscillates more across groups (higher std). We can accept the hypothesis formulated initially. It is possible to infer when a group of people is in a collaborative situation laying on the application and audio traces, taking into consideration the limitations of each algorithm. Even when our model was limited to quantitative data we could get enough information to infer if the group of learners were potentially engaged in collaborative interactions.

Table 4. Results of the leave one out cross validation.

	Log sets of 30 seconds				Log sets of 60 seconds				Log sets of 90 seconds			
	BNet	NB	C4.5	**BFT**	BNet	NB	C4.5	BFT	BNet	NB	C4.5	BFT
Accuracy	.667	.688	.661	**.645**	0.660	0.694	0.656	0.673	0.648	0.642	0.605	0.606
Standard deviation	.099	.119	.140	**.135**	0.178	0.179	0.199	0.167	0.172	0.163	0.194	0.167
Correlation(C)	-0.095	-0.276	-0.206	**-0.246**	0.029	-0.143	-0.101	0.160	-0.324	-0.090	-0.298	-0.193
Correlation (SC)	**-0.61**	**-0.79**	**-0.343**	**-0.382**	**-0.623**	**-0.688**	**-0.423**	**-0.301**	**-0.695**	**-0.550**	**-0.775**	**-0.553**
Correlation (NC)	0.413	0.649	0.34	**0.391**	0.334	0.491	0.312	0.050	0.623	0.401	0.650	0.450

6 Conclusions

We presented an overview of our work to infer the extent of collaboration within groups of learners building on the foundation of collaborative learning theories [10] and data mining techniques. Our aim is to explore the intersection between the quantitative traces of peers' interactions and the research area of collaborative learning. Our approach does not take into account groups' performance. Indeed, we did not find any relationship between collaboration and this feature, obtaining a correlation of -0.052. We found that the main indicators of collaboration are the quantity and heterogeneity of verbal participation. However, the quantitative data does not tell the whole story of a group. The performance of our classifier is good enough to provide valuable information which is currently not automatically available. It would enable a teacher to see if an activity that was intended to be collaborative really was so. It would also give teachers and learners a good indication of how well each group was collaborating. The preliminary results of this study are promising and further research must be done to assess if they apply to other domains.

References

1. Stahl, G.: Collaborative learning through practices of group cognition, in Proc.CSCL 2009, International Society of the Learning Sciences: Rhodes, Greece. pp. 33--42 (2009)
2. Soller, A., Martinez A., Jermann P. and Muehlenbrock M.: From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. JAIED. 15(4): pp. 261--290 (2005)
3. Hattie, J. and Timperley H.: The Power of Feedback. Review of Educational Research, 77(1): pp. 81--112. (2007)
4. Chaudhuri, S., Kumar R., Howley I. and Rose C.: Engaging Collaborative Learners with Helping Agents, in Proc. AIED 2009, IOS Press. pp. 365--372 (2009)
5. Duque, R. and C. Bravo: A Method to Classify Collaboration in CSCL Systems, in Adaptive and Natural Computing Algorithms, Springer Berlin / Heidelberg. pp. 649--656 (2007)
6. Perera, D., Kay, J., Koprinsca I., Yacef K. and Zaiane O.: Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. IEEE TKDE 2009. 21: pp. 759--772 (2009)
7. Anaya, A. and J. Boticario, Application of machine learning techniques to analyse student interactions and improve the collaboration process. J. Expert Systems with Applications. 38(2): pp. 1171--1181. (2011)
8. Wallace, J., Scott, S., Stutz, T., Enns, T. and Inkpen, K.: Investigating teamwork and taskwork in single-and multi-display groupware systems. Personal and Ubiquitous Computing 13(8), pp. 569--581 (2009)
9. van Dongen, B.F., de Medeiros, A. K., Verbeek, H., Weijters, A. and ,van der Aalst, W. M.: The ProM Framework: A New Era in Process Mining Tool Support. pp. 444--454. (2005)
10. Dillenbourg, P.: What do you mean by 'collaborative learning'?, in Collaborative Learning: Cognitive and Computational Approaches. Elsevier Science. pp. 1--19. (1998)
11. Weinberger, A. and F. Fischer,: A framework to analyze argumentative knowledge construction in CSCL. Computers & Education. 46(1): pp. 71--95. (2006)
12. Harris, A., Rick J., Bonnett V., Yuill N., Fleck R., Marshall P. and Rogers Y.: Around the table: are multiple-touch surfaces better than single-touch for children's collaborative interactions? in Proc. CSCL. 2009, Rhodes, Greece. pp. 335--344. (2009)
13. Bousbia, N., Labat J., Balla A. and Rebai I.: Analyzing Learning Styles using Behavioral Indicators in Web based Learning Environments, in Proc. EDM. 2010. (2010)