Visualizing Large Quantities of Educational Datamining Information

Sandra Gama, Daniel Gonçalves
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa
sandra.gama@ist.utl.pt, daniel.goncalves@inesc-id.pt

Abstract

Providing the educational community with tools to analyze educational processes may result in a more effective education. Applying Data Mining techniques to educational data results in information on educational settings which, however, comprehend an extensive set of symbolic patterns that are usually difficult to understand. Visualization, due to its potential to display large quantities of data, may overcome this limitation. We used the results of educational data mining techniques that had been applied to analyze the interdependence among courses in a university program and studied visualization mechanisms to enable the analysis of such patterns. We created a multi-level visualization, in which each level depicts a semester with corresponding courses. We have studied visual connectors to display a high number of interrelations between courses. User tests have shown the effectiveness of a connector which combines visual merging techniques with Bezier curves to represent course interrelation.

Keywords— **Educational Information Visualization, Data Visualization, User Interfaces**

1 Introduction

The number of students in both traditional and online education has grown considerably over the last decades. University students' numbers have increased considerably in a global scale. In fact, the global enrollment rate increased from 8.5%, in 1970, to 24.7%, in 2006[4]. Concerning online education, the creation and profusion of free MOOC (Massive Open Online) courses in which students from all over the world may participate also contributed to a growth in the number of students worldwide. As a result, CMS (Course Management Systems) and LMS (Learning Managament Systems) became popular and had a great impact in the boost of distance education [5].

With the growing number of students in both traditional education and online courses, a very large set of data emerges from students' curricula. This information, if explored effectively, may be crucial to improve education processes.

The application of data mining techniques to educational information is an emergent research topic, providing the means to analyze data from educational settings, ranging from student behavior to teaching strategies and program coordination. EDM (Educational Data Mining) is an emerging discipline with the goal of applying data mining techniques to data from educational settings. It provides relevant patterns based on available course information and makes it possible to make predictions based on educational data provided. However, EDM results consist of a large set of textual patterns which are difficult to understand due to their visual complexity. Since information interpretation is of utmost importance for EDM to be useful and effective, this limitation must be overcome. In order to overcome this obstacle, the user must be involved in the exploration process in a way that combines creativity, flexibility and general knowledge [6]. Visualization has the potential to overcome this challenge: it an excellent means to display large quantities of data and alleviates cognitive load associated with information interpretation [12]. Displaying the results of educational data mining techniques in a meaningful way will make it enable study program coordinators and professors to be aware of problems that would otherwise remain unnoticed.

We created a visualization that displays educational patterns in a way that makes them easy to navigate and interrelate, allowing an effective analysis. The information to visualize is the result of EDM techniques that had previously been applied to analyze the interdependence of success among courses in an university program in the context of a research project, Educare [2].

We created a multi-level layered visualization for representing a study program's set of courses, focusing on the effective representation of interrelations among them. We have studied two mechanisms to represent such interrelations: a simple line connector and a technique which combines visual merging mechanisms with Bezier curves.

In section 2 we analyze and discuss relevant work in this area. We then introduce the EDM patterns. Section 4 describes our visualization, focusing on course interrelation. In section 5 we present and analyze the results from evaluating interrelation visualization techniques. Lastly, we draw several conclusions and guidelines for future work.

2 Related work

Several tools have been created to represent educational information. Taking online education into account, CMS allow the creation of virtual classrooms, making it possible to remotely participate in discussions and manage classes, generating large amounts of data that need to be managed in a way that provides teachers relevant information. In order to overcome this challenge, CourseVis has been created [8], which is used as an extension to CMS that allows interactive data exploration and manipulation through different visualization mechanisms: (i) the representation of students participating in a forum, in which threads are represented as spheres with size proportional to the number of students involved; (ii) the Cognitive Matrix, which consists of a matrix in which students' names and their knowledge regarding course concepts; (iii) a matrix-like visualization tool, in which students' behavior is depicted, highlighting content access, attendance and progress. However, users found some aspects unintuitive, such as the color arrangement and visual overlapping [9]. Attempting at overcoming CourseVis' limitations, GISMO (Graphical Interactive Student Monitoring System) [7] was created. It represents LMS information, which is quite complex and difficult to read. Focusing on the previous system's two-dimensional behavior visualization, GISMO allows interactive exploration of access and resource details, and provides the means for the exploration of students' behaviors that were previously considered relevant. After using GISMO on an online course, it seemed effective for monitoring class and individual behavior, evaluating participations in forums and enhancing the course.

Concerning traditional education, results from evaluation processes may be difficult to read and interpret and patterns finding is nearly impracticable in this context. Several research has been done in this context. One is AVOJ (Analysis and Visualization for Online Judge of Teaching Data) [14]. This tool allows the exploration and visualization of data regarding students' performance, providing the means for grouping students according to their grades and other aspects such as study habits. Another interesting study is that of Xiaoya et al. [15], who created a visualization for analysis of college students' scores in an english course. They followed the parallel coordinate approach, in which N equidistance parallel axes are used to represent dimensions of a multidimensional dataset. However, since such a visualization is not sufficient to find out further information, several models have been used which improve result reliability. These models, besides allowing the representation and interactive manipulation of data, allow an immediate information overview, enabling a more efficient data analysis. Trimm et al. [11] have created a visualization in which students are grouped according to their grades as well. Groups may be viewed using compositions that show their features and variations in time. In order to use composition, information on each student's history consists of a two-dimensional trajectory, represented in two axes, allowing an appropriate time-centered representation of data features. In order to understand why a significative number of students drop their computer science study programs, a visualization tool has been created which allows the visualization of repeating patterns on student failure and success while attending the program [13]. Due to the great amount of information available, a colored node and edge structure has been used. Colors represent students' performance, allowing the differentiation of student groups and identifying groups with similar behavior. The visualization is interactive and allows the selection of student categories, such as *students who repeat at least a course* or *students who never failed a course*, making it possible for teachers to draw a set of conclusions regarding repetitive course failures or implications of a given course's grade on the success at other courses.

All aforementioned approaches focus on visualizing information on education processes, either traditional or online. They present different techniques and methodologies for visualizing such data in their particular context. Given the scope of our study and the particularities of our data, in which a visualization that can interrelate courses and show their interdependences presents more relevance, we highlight the work by Wortman and Rheingans' study [13] which, although providing interaction mechanisms, does not allow either comparison between courses regarding large quantities of data or specific pattern analysis. We have, thus, studied a technique to represent interrelations among study program courses while attempting to address the shortcomings of existing solutions.

3 Educational Patterns

One of the problems in using educational data in a way that improves success is the acquisition of background knowledge, both on current teaching strategies and students' frequent behavior. We used the result of sequential pattern mining that had previously been applied to data that had been gathered for nine years on an undergraduate study program on information technology and computer science [2]. The goal of sequential pattern mining, given a set of sequences and some user-specified minimum support threshold, is to discover the set of sequences that are contained in at least σ sequences in the dataset, that is, the set of frequent sequences [1]. This technique allows for the discovery of frequent sequential patterns among the recorded behaviors that are consistent with existing background knowledge. Such knowledge may be represented by a context-free language, which plays the role of a con-

straint in the sequential pattern mining process. Not only does this method gather expected patterns based on background knowledge but the use of constraint relaxations also enables the discovery of patterns that correspond to deviations to the expected behavior, making some potentially relevant trends evident that were previously unknown [2]. In order to do so, the curriculum knowledge has been represented as a finite automaton, which establishes the order of subjects students should attend to finish their graduation and provides information grouped by semester. Sequential pattern mining with three different support threshold values (50%, 25% and 20%) has been performed, resulting in three different sets of patterns. Evidently, the lower the support threshold values, the higher the number of patterns result from applying sequential pattern mining. As a result of applying the aforementioned methodology, a number of textual patterns have been generated. Such patterns observe the following structure:

$$Pattern_i = (semester_1, ..., semester_N, total_{students}),$$

$$semester_j = course_1 \vee (course_1, ..., course_M)$$

Some examples are:

- $(fex, 2000)$: 2000 students completed *fex* on the first semester;
- $(fex, tc, 1400)$: 1400 students completed *fex* on the first semester and *tc* on the second semester;
- $((fex, am1), tc, 1000)$: 1000 students completed *fex* and *am1* on the first semester and *tc* on the second semester;
- $((fex, am1), (fisica1, tc), am2, 800)$: 800 students completed *fex* and *am1* on the first semester followed by *fisica1* and *tc* on the second semester and *am2* on the third semester.

Even though textual information makes it difficult to understand particular patterns and provides little insight into general trends, this pattern structure provides us with the means to gather information on course interrelations among different semesters that an effective visualization will be able to make evident.

4 Visualizing Large Quantities of Patterns

We created a visualization which consists of a multilevel interactive layered representation of semesters where relationships among courses are depicted (Figure 1).

Courses are displayed as circles with size proportional to the total number of students who completed them: the higher the number of students, the bigger the circle radius. Whenever EDM patterns provide information regarding course failure, the course circle is sub-divided into two semicircles displaying information both on course approval and failure through western conventional positive-negative color coding [12]. The leftmost, green, semicircle represents the number of approved students and the rightmost, red, semicircle, displays the number of students who were not able to finish the course successfully. Similarly to courses with only information on success, semicircle size is proportional to the data being represented, providing immediate information and comparison between success and failure rates among courses without requiring further data exploration.

By selecting a course, we may see all the courses with which this one has any type of interrelation through visual connectors. Connector thickness is proportional to the number of students who verify the pattern: the thicker the line, the more students participate in the current pattern. Color is assigned to each pattern individually, in order to avoid visual confusion and allowing immediate line discrimination. We did not use fully saturated colors, in order to keep our visual artifacts from competing for the user's visual attention [12]. When the course circle is no longer selected, this information is hidden.

4.1 Visual course connectors

We created two different types of connectors for representing interrelations among courses (patterns). The first consists of simple line connectors, as depicted in Figure 2.

This line connector, however, generates some clutter when more complex patterns are represented. We did, however, realize that several patterns are repeated (e.g., students who completed both course A and course B did complete course C, but they also completed course D). Instead of representing repeated patterns separately, we considered creating merging mechanisms to simplify our representation of interrelations among courses.

However, merging lines did not prove successful in eliminating clutter, hence we represented course connections through cubic Bezier curves. Such connector is depicted in Figure 3.

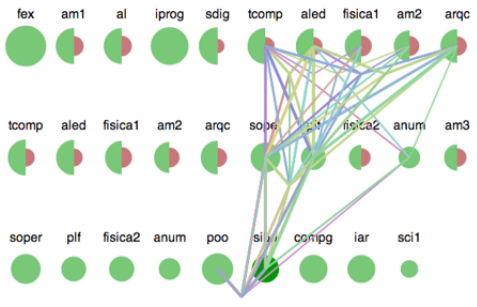

Figure 2: Line connectors

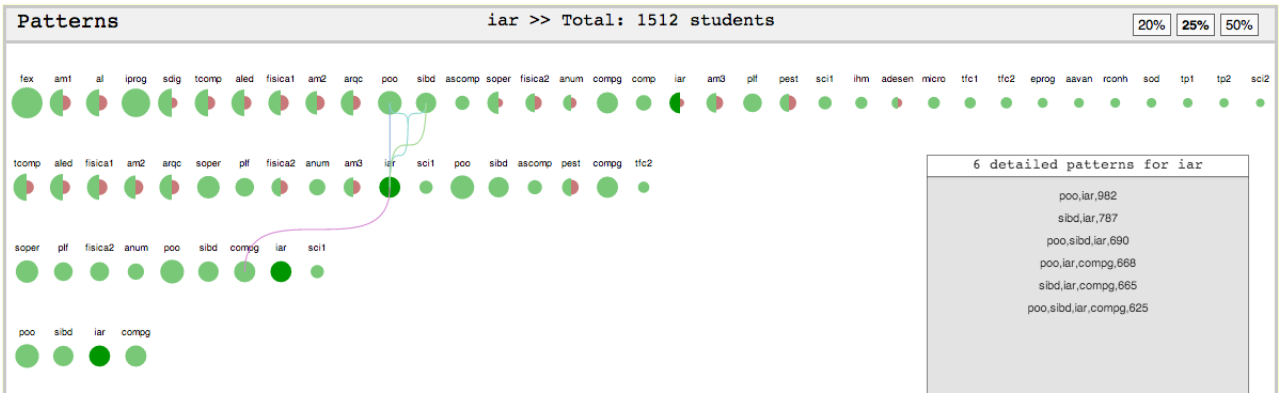

Figure 1: Multilayered visualization depicting courses as circles, arranged by layers (course semesters). Visual curve connectors represent interrelations among selected courses.

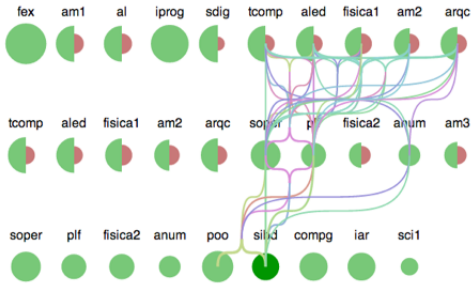

Figure 3: Merged curves' connectors

Since line thickness represents the number of students who completed a particular set of courses (described in the EDM pattern), by merging course sequences we added line thickness when merging repeated patterns in order to keep context and coherence.

5 Evaluation

We performed a user study in order to evaluate our method for grouping and merging repeated educational patterns, compared to binary connectors. As a result, taking the aforementioned context of our study into account, we intend to infer: (i) the effectiveness and efficiency of our solution, (ii) whether our method provides good usability and learnability and (iii) how satisfied subjects feel after task performance.

5.1 Study protocol

Two different test settings were created, corresponding to: *S1: Line connectors*; *S2: Merged Bezier curve connectors*. We created a number of scenarios corresponding to sets of patterns with increasing size to apply to each test setting. We took into account the following courses: (1) *IAR*: 8 patterns; (2) *PO*: 22 patterns; (3) *fisica1*: 50 patterns; (4) *ALED*: 103 patterns; (5) *AL*: 1013 patterns. For each scenario, subjects were asked to complete two tasks. Results on task performance will allow us to compare the effectiveness of either one of the visual mechanisms we are studying. Tasks were: (i) *Name the courses which have any type of interrelation with the current course.* and (ii) *Name the three most relevant patterns regarding the current course.*

We conducted a study with 15 users. Out of our participants, we had 10 (66.67%) male subjects and 5 (33.33%) female. Their ages range between 18 and 24 ($\overline{X} = 20.67, \sigma = 1.54$). Out of our test subjects, 6 (40.00%) have a BsC degrees and 9 (60.00%) have completed high school.

Participants were given a small verbal introduction to the study with a brief description of the context and the main visualization tools. Then, for each scenario (line connectors or merged curves), we followed the same protocol: subjects were presented with the current course interrelation visualization mechanism and asked to perform both tasks for each scenario (course). After completing the tasks, they were allowed to freely interact with the visualization while making verbal comments. Participants were then asked to complete a small satisfaction questionnaire consisting of the SUS (System Usability Scale) [3]. Tests were conducted in a well illuminated room and all subjects were provided the same laptop for performing their tasks. We measured the time subjects took to complete every task.

5.2 Results

Measured time to complete each task is summarized in Tables 1 and 2, taking course scenarios and methods into account. In general terms, it is clear that users took less

time to complete tasks using the merged curves' method.

Scenario	Method	Average	St. deviation
1	Lines	12.27	2.66
	Merged curves	10.60	2.32
2	Lines	16.13	3.98
	Merged curves	16.07	3.33
3	Lines	16.93	5.02
	Merged curves	17.00	4.33
4	Lines	44.67	9.54
	Merged curves	25.53	4.29
5	Lines	51.67	8.20
	Merged curves	36.00	3.61
Average	Lines	28.33	3.49
	Merged curves	21.04	1.74

Table 1: Average performance time for each visualization method (Task 1).

Scenario	Method	Average	St. deviation
1	Lines	46.87	7.71
	Merged curves	9.33	3.15
2	Lines	39.93	12.50
	Merged curves	12.40	3.87
3	Lines	41.53	7.39
	Merged curves	15.00	3.36
4	Lines	19.07	7.08
	Merged curves	21.87	4.07
5	Lines	62.20	16.97
	Merged curves	18.60	3.66
Average	Lines	41.92	5.97
	Merged curves	15.44	1.28

Table 2: Average performance time for each visualization method (Task 2).

In order to verify these results, we performed further statistical analysis. We run Shapiro Wilk test, having found evidence against normality in some of our data Hence, we applied Wilcoxon Signed-Rank Tests to each task and scenario, comparing the two methods of representing visual connections. Regarding Task 1, even though results are not statistically significant for less complex patterns ($T_S1 = -1.91, p = 0.06$, $T_S2 = -0.20, p = 0.84$, $T_S3 = -0.26, p = 0.79$), when the number of patterns increases, results are significantly better for the merged curves' method ($T_S4 = -3.41, p < 0.01, r = -0.62$, $T_S5 = -3.29, p < 0.01, r = -0.60$), as it is in general terms ($T_AVG = -3.41, p < 0.01, r = -0.62$). As for Task 2, results are not statistically significant for one single scenario ($T_S4 = -1.04, p = 0.3$), whereas the merged curves' approach proved significantly better ($T_S1 = -3.41, p < 0.01, r = -0.62$, $T_S2 = -3.41, p < 0.01, r = -0.62$, $T_S3 = -3.41, p < 0.01, r = -0.62$, $T_S5 = -3.41, p < 0.01, r = -0.62$), as it is in general terms ($T_AVG = -3.41, p < 0.01, r = -0.62$).

Regarding the number of correct answers (scored from 0 to 5), the merged curves' method ($\overline{X}_{Task1Curves} = 4.87$, $\sigma_{Task1Curves} = 0.35$, $\overline{X}_{Task2Curves} = 4.80$, $\sigma_{Task1Curves} = 0.56$)) generally yielded better results than the lines' method ($\overline{X}_{Task1Lines} = 3.80$, $\sigma_{Task1Lines} = 1.32$, $\overline{X}_{Task1Lines} = 0.67$, $\sigma_{Task1Lines} = 0.98$). A further statistical analysis was applied to verify such results. Shapiro-Wilk tests showed evidence against normality ($T_{T1L} = 0.77, p < 0.01$, $T_{T1C} = 0.41, p < 0.01$, $T_{T2L} = 0.73, p < 0.01$, $T_{T2C} = 0.42, p < 0.01$). Hence, we applied Wilcoxon Signed-Rank Tests, which have shown statistical evidence which supports our findings ($T_T1 = -2.52, p < 0.05, r = -0.46$, $T_T2 = -3.40, p < 0.05, r = -0.62$).

Taking the final satisfaction questionnaire into account, calculating the SUS score [10], we obtained a score of 68.00 points for the lines' approach and 84.5 for the merged curves' method. While the first did not yield particularly encouraging results, the usability of the second method was placed at the top 10% percentile with a very good evaluation at both usability and learnability.

5.3 Discussion

Evaluation has shown that the implementation of a method which combines repeated patterns and visually represents such patterns as merged curves has proven promising. In fact, success rates regarding the accuracy of answers provided for each task are nearly 100%, showing that users were able to obtain the required information to complete their tasks. Differences are particularly relevant when large amounts of data are represented, showing that merging repeated items in such a way alleviates clutter and makes general data more easy to read and understand. Finally, user satisfaction questionnaires reiterated our results, demonstrate the effectiveness of our method for visualizing large amounts of patterns in the particular context of our study.

Conclusions

With the increasing number of students in traditional and online education, a large amount of data emerges from learning activities. If analyzed properly, it may provide the means for refining education processes. In this context, data mining techniques have proven effective in finding relevant patterns. However, the result of applying such techniques often leads to a complex set of data which is difficult to read, interpret and analyze. Overcoming this limitation will provide the means to generally perceive this information as a consistent whole and realize particular data

aspects as well. We attempted to do so, by relying on visualization mechanisms, which are outstanding at displaying great volumes of data while allowing efficient and effective data interpretation.

We created a visualization that combines different interaction mechanisms. It revolves over a central representation of layers that correspond to course semesters. On each semester layer, corresponding courses are depicted as interactive circles: when one is selected, patterns are depicted as line connectors which associate the current course with the ones with which it has a simultaneity or precedence relationship. We have studied an effective way to represent course interrelation through the creation and evaluation of both line connectors and merged Bezier curves.

User tests have shown that the Bezier curve connectors are promising for representing large quantities of data regarding relationships among courses. Participants were able to immediately perceive several interrelations among courses and analyze such interrelations. Furthermore, they were able to learn and easily use the system, proving the potential of our visualization method for representing large amounts of data in a perceivable way. As a result, we expect our contribution to provide the educational community with patterns that were not evident otherwise. As a result, program coordinators and professors will have the means to have further insight into courses and programs and will be able to track down particular problems, which will lead to increased success.

Acknowledgements

This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project Educare - PTDC/EIA-EIA/110058/2009 and INESC-ID multiannual funding - PEst-OE/EEI/LA0021/2013.

References

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the IEEE International Conference on Data Engineering*, pages 3–14, 1995.

[2] C. Antunes. Acquiring background knowledge for intelligent tutoring systems. In *Proceedings of the International Conference on Educational Data Mining*, pages 18–27, 2008.

[3] J. Brooke. Sus: A quick and dirty usability scale, 1996.

[4] Y. Gao. A study on mass higher education in the world-based on comparative perspectives. In *International Conference on Education and Management Technology (ICEMT)*, pages 528–530, 2010.

[5] J. Kay, P. Reimann, E. Diebold, and B. Kummerfeld. Moocs: So many learners, so much potential ... *IEEE Intelligent Systems*, 28(3):70–77, 2013.

[6] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[7] R. Mazza and L. Botturi. Monitoring an online course with the gismo tool: A case study. *Journal of Interactive Learning Research*, 18(2):251–265, April 2007.

[8] R. Mazza and V. Dimitrova. Generation of graphical representations of student tracking data in course management systems. In *Ninth International Conference on Information Visualisation, 2005. Proceedings*, pages 253–258, 2005.

[9] J. Steele N. Iliinsky. *Designing Data Visualizations*. O'Reilly, 2011.

[10] J. Sauro. *A Practical Guide to the System Usability Scale: Background, Benchmarks and Best Practices*. CreateSpace, 2011.

[11] D. Trimm, P. Rheingans, and M. desJardins. Visualizing student histories using clustering and composition. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2809–2818, 2012.

[12] C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2012.

[13] D. Wortman and P. Rheingans. Visualizing trends in student performance across computer science courses. In *Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education*, pages 430–434, 2007.

[14] W. Xiaohuan, Y. Guodong, W. Huan, and H. Wei. Visual exploration for time series data using multivariate analysis method. In *8th International Conference on Computer Science Education (ICCSE)*, pages 1189–1193, 2013.

[15] G. Xiaoya, L. Kan, and L. Ping. Visual analysis of college students' scores in english test. In *4th International Conference on Computer Science Education (ICCSE)*, pages 1816–1819, 2009.