International Journal of Creative Interfaces and Computer Graphics
Volume 7 • Issue 1 • January-June 2016

Visualizing Sequential Educational Datamining Patterns

Vilma Rodrigues Jordão, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
Sandra Gama, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
Daniel Gonçalves, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

ABSTRACT

The use of educational datamining techniques to elicit patterns in student behavior and learning outcomes can be a useful for the analysis of the effectiveness of teaching strategies and the coordination of study programs. However, results from those techniques are, often, large sets of symbolic patterns, numbering in the thousands, usually presented in text format. This makes them hard to understand which, coupled with the lack of an overall view, hinders a more comprehensive data analysis. The authors propose that information visualization techniques can be used to display relevant information in those patterns in effective ways, allowing decision makers to better insights about the reality at hand. They present a solution built upon two linked views, one based on node-link representations and another a multi-matrix representation. The complementarity of both visualization techniques allows the most important patterns to be immediately apparent, while at the same time permitting their interactive exploration in meaningful ways. The authors performed user tests proving their effectiveness.

KEYWORDS

Color Blending, Education, Educational Datamining, Human-Computer Interaction, Information Visualization, Sequential Patterns

INTRODUCTION

The number of students in both traditional and online teaching has grown considerably over the last decades. At the college level, there has been an increase from 8.5% (in 1970) to 24.7% (in 2006) in the percentage of people that have at least one degree (Gao, 2010). Online learning in particular has seen a huge boost given the emergence of Massive Open Online Courses (MOOCs), in which people from all over the world can enroll and participate. As a result, Course Management Systems (CMS) and Learning Management Systems (LMS) have become very popular and have now a large impact on distance learning (Kay, 2013).

To understand the effectiveness of a learning program or a particular course, a number of indicators must be considered, the not least important of which being student success. This in itself includes several factors such as drop-out rates and final grades but is, nevertheless, a simple indicator of how well things are going and can provide early warning signs for systemic problems that must be corrected. Also important are the inter-relations between courses in a same degree. Understanding them will give us insights into critical paths and bottlenecks in a curriculum. All of this can be used to fine tune curricula, leading to an increase in teaching quality, better learning outcomes and student satisfaction. Small student/course samples are not enough to elicit any meaningful, generalizable

DOI: 10.4018/IJCICG.2016010101

information. However, given the aforementioned increase in numbers, coupled with the development of increasingly sophisticated LMS and the school's own information systems, a wide and encompassing range of data is now within reach. Its timely analysis can be crucial for the improvement of the teaching-learning process.

Alas, the available data is seldom in a format that is amenable to analysis and exploration. Of real help for decision makers would be *meaningful patterns* that can be elicited from that data, rather than having to go through thousands of individual records. Only in that way can the overall picture be known and understood. The application of data mining techniques in this context is an emerging research field. It provides the means to analyze educational data, from student behaviors to teaching strategies and course coordination. For this domain, it takes the moniker of Educational Data Mining (Romero, 2010). EDM provides relevant patterns based on the available data and is a tool both for the study of what went on in the past and to make informed predictions for the future. However, the results from an EDM process usually consist of extensive sets of behaviors, described in the form of technical patterns that are represented textually or symbolically. While much more meaningful than the original data, the amount of patterns can still be daunting and not conducing to an encompassing analysis. Furthermore, the understanding of the patterns themselves often requires a reasonable knowledge of the underlying data mining algorithms and statistical models, which many analysts probably do not possess.

Globally, it is important to provide an overall view of all patterns as a coherent whole, allowing the identification of commonalities and differences. It is also necessary to interpret individual patterns and establish relations between them. Furthermore, a tool that allows this should involve the users and allow them to take advantage of their creativity, flexibility and domain-specific knowledge (Keim, 2002). One possible approach to this problem is to use Information Visualization. One of its aims is to help understand large amounts of data by leveraging on the capacity of the human visual system to discover trends, patterns and outliers (Heer, 2010). Furthermore, a well-designed visualization can effectively represent large amounts of data and alleviate the cognitive load associated with interpreting it (Ware 2004). A visualization of the results from EDM has the potential to provide the insights it needs.

In the context of the Educare project, the authors had access to sets of EDM patterns (Antunes, 2008). Their goal was the analysis of the interdependencies of college courses within an undergraduate degree. They are based on the number of students that pass or fail each course, and the sequence in which they do so. An example of a pattern (stated in natural language) is: *50% of all students that failed course A and passed course B on a given semester were not able to pass course C on the next semester they enrolled*. This example illustrates the main challenge in creating a visualization for EDM patterns, besides their large numbers: their sequential nature. These patterns are structured as sequences of occurrences through time (in this case, school semesters). This sequential nature is crucial for their understanding and must be central in the visualization. However, most work in the visualization of data mining patterns focuses only on static patterns, representing a single instant of time.

In this paper the authors describe how a visualization, EduVis, was created, which allows the representation of sequential EDM patterns and their exploration using two complementary techniques: a multi-layer link-node visualization that makes dependencies evident and a multi-matrix-based visualization supported on heatmaps to highlight overall pass/fail trends. Both take advantage of color blending techniques to present more degrees of freedom without unduly overcomplicating the visualization in spatial terms. A set of user tests has validated their understandability, correctness and usefulness.

In the next section the related work in the area of the visualization of educational information will be described. Next, the nature of the patterns to visualize will be discussed. Following that, EduVis will be presented, its design rationale and underlying decisions, followed by its evaluation, supported on user studies, after which the authors will draw some conclusions, pointing to possible directions for future work.

RELATED WORK

With the growing number of students in both traditional and online education, several tools for the graphical representation of educational information have been created.

Taking online education into account, LMS allow the creation of virtual classrooms, making it possible to participate in remote discussions and class management, generating a large amount of data that need to be managed to provide instructors with relevant information about student performance. To address this challenge, CourseVis was created (Mazza, 2005). It works as an extension to the LMS that allows the interactive exploration and manipulation of data through different visualization techniques such as a tridimensional representation, where different topics are represented as spheres with a size proportional to the number of participants. Another technique is the Cognitive Matrix, a student-concept matrix in which intersections are colored in a scale ranging from green (pass) to red (failure). A third tool represents student behavior through sorted graphs and text, such as attendance and learning progress. A user study has shown these visualization techniques to be effective in delivering student profiles. However, users had problems in understanding the information due to the overlap of several graphical elements (including but not limited to the 3D view).

The limitations found in CourseVis led to the creation of the Graphical Interactive System for Monitoring Students (GISMO), which represents data from a LMS that would otherwise be complex and hard to understand (Mazza, 2007). GISMO is integrated into the Moodle LMS but could be adapted for other platforms. It focuses on a two-dimensional visualization and allows the exploration of student behaviors found interesting in CourseVis. After a trial usage for an online course, GISMO proved effective in facilitating the understanding of student behavior as well as the effectiveness of the different evaluation moments. This allowed the redesign of the course according to the students' actual needs.

Traditional teaching is strongly based on evaluation processes such as exam grades, attendance and participation, some of which are complex and hard to interpret. In this regard, AVOJ (Xiaohuan, 2013) aims at highlighting the different capabilities of students by visualizing performance data, grouping students according to grades and similar factors. However, it uses a color scheme with maximum brightness and saturation values that is known to become overwhelming to some users and make it more difficult to find patterns (Iliinsky, 2011). The system allows direct comparison of general trends, based on bar charts displaying real-time statistics and making it possible to understand some specific things such as the ways students manage study sessions.

Another interesting study is that of Xiaoya et al (Xiaoya, 2009), a visualization of the results of college students of an English course, using parallel coordinates where N axis represent dimensions in a multidimensional dataset. It allows for: (i) classification, where the user can divide the data into sets corresponding to different lectures; (ii) averaging, that depicts, by interacting with the main chart, average values for each set; (iii) boxplots, to assess data dispersion; (iv) axis permutation, changing their order to find new insights; (v) correlation of two data sets along the different axis; (vi) association, where the value of attribute A is estimated based on attribute B; and (vii) roll-up and drill, allowing the hierarchical representation of data. These features, going beyond the interactive exploration of data, allow the visualization of a large number of variables simultaneously, providing a holistic approach.

Trimm et al. (Trimm, 2012) have created a visualization where students are grouped according to their grades. These groups can be visualized in a way that allows the display of their evolution through time. To that end, the history of each student's performance is shown as a two-dimensional trajectory along two axis. As such, this approach is not really suited to simultaneously display features and tendencies of large student sets. Even so, a color blending technique, *color weaving*, merges trajectories from different students to make clear the overall trend and dispersion in the data. In practice, this makes accurate assessments hard. User tests have shown, nevertheless, that after visual inspection users can find relevant patterns in the data.

To understand student drop-out rates in a Computational Science degree, Wortman created a tool to visualize success and failure patterns (Wortman, 2007). A node-like approach was used. Nodes represent events and their size is proportional to the number of students that participated in it. Edges represent student's trajectories along those events. Their thickness is proportional to the number of students. Color represents student performance and allows the visual identification of student groups according to that measure. The visualization makes it possible to select students of a similar profile using criteria such as *"students that never failed course X"*. As a result, professors could gain insights about the causes for high drop-out rates. However, the cluttering arising from the overlapping of links in the visualization made a deeper analysis impossible.

All the aforementioned approaches focus on the visualization of traditional and online teaching processes, using different techniques and approaches. Given the goal of displaying EDM patterns, research by Wortman (Wortman, 2007) and Trimm (Trimm, 2012) are the most relevant. However, the interactive exploration allowed by Trimm is limited. Wortman's solution supports such an exploration, but it does not allow the comparison between courses nor the highlighting of specific patterns. Indeed, none of the discussed works focus on the visualization of patterns but, rather, on the original data, leaving the patterns to be informally discovered by users using visual inspection. To bridge that gap, a visualization where patterns are central was created.

EDUCATIONAL PATTERNS

With the increasing number of students in higher education, and in order to provide students with study programs that meet their needs, it is necessary to identify existing patterns regarding success and failure in their curricula. As a result, it will be possible for the education community to acquire knowledge on current problems and challenges and act in order to rectify such limitations and thus improve teaching processes.

In this work, data were collected during 9 years, regarding the performance of students in the first three academic years of an IT and Computer Engineering undergraduate degree. Educational Data Mining (EDM) techniques were applied to the data, particularly sequential data mining (Agrawal 1995). Besides meeting the expected standards based on background knowledge, this method allows the discovery of patterns corresponding to deviations from the expected behavior, hence making evident relevant trends that were previously unknown (Antunes, 2008). Using this technique, it was possible to extract patterns that describe the number of students who passed and failed certain courses, as well as relations of precedence and simultaneity. An example of precedence would be, "90 students who failed course A in one semester, also failed course B in the following semester." A simultaneity relationship would be "10 of the students who passed course C in a given semester, also passed course D in the same semester."

Thus, knowledge on the study program was represented as a finite automaton, establishing the order in which courses must be completed for students to succeed in the study program. Data mining took three different support threshold values into account (50%, 25% and 20%, corresponding to the percentage of the total number of students to which the pattern is to be applied), resulting in three different sets of textual patterns, grouped by semester. These correspond to enrollment semesters and not the semesters in the study program. That is, in the examples below, when courses are referred to as part of the "first semester" or "second semester", the authors are referring to consecutive semesters of students' enrollment, not academic year.

Patterns refer to the number of students who either passed or failed a set of courses. They are sequences, and so start with '[' and end with ']'. They are organized by semester, in which content consists of the subjects of each semester; hence they observe the following structure:

$\text{pattern}_i = (\text{semester}_1,\ldots,\text{semester}_N\ \text{total}_{\text{students}}),$
$\text{semester}_j = \text{subject}_1 \vee (\text{subject}_1,\ldots,\ \text{subject}_N);$

For instance:

['am2', '1683']: 1683 students passed course *am2*.

['fex', ['tcomp', 'aled'], '1168']: 1168 students passed *fex* in the first semester, as well as *tcomp* and *aled* in the second semester.

['fex', ['~arqc', '~fisica1'], '591']: 591 students passed *fex* in the first semester but failed both *arqc* and *fisica1* in the second semester. In this last case, failing a course is represented by the symbol '~'.

Although the textual format makes it difficult to understand specific patterns and provides limited overall information, this structure provides information about the relationships between semesters on a study program that an effective visualization will be able to highlight. Another problem is the amount of patterns. For 25% and 20% support, they number in the thousands, for this particular dataset.

PROPOSED SOLUTION: EDUVIS

The generated educational patterns that the authors aimed at visualizing, besides their large number (especially considering lower support thresholds), have several dimensions (semesters, courses, success, failure, etc.). In addition, as described above, patterns are sequential, reflecting a temporal nature: part of the events takes place simultaneously and part over time. The temporal dimension is very rich, allowing the depiction of typical students' paths over time. However, viewing precedence and simultaneity relations was a considerable challenge since, traditionally, most visualization techniques are applied to either static or dynamic data but not both.

In fact, it was important to visualize patterns that can consist of an arborescent form if viewed as a whole, for all the students. In its original form, patterns appear as sequences of events with associated probabilities, making it difficult to use them for analysis and as a decision support tool. An effective display would correct this aspect. A survey of possible alternatives was conducted to temporarily represent the available information.

An adapted version of the classic Minard display on the Napoleonic wars (Figure 1) was initially considered, which seemed a natural starting point. It consists of a view over time which highlights certain moments in a timeline, consistent with the curricular path throughout the semesters. Furthermore, this representation enables a natural implementation for success (troops remain in the main group) and failure (troops who separate or perish).

The creation of various paper prototypes led the researchers to conclude that the use of parallel sets sharing some characteristics with the desired view would be suitable, having thus implemented a working version of a view based on this approach (Figure 2).

However, tests with real data showed that this solution was insufficient. In fact, one of its biggest advantages would be the easy representation of converging branches, corresponding to patterns, which, through different paths, led to the same outcome. However, this information is not liable to be faithfully representative of existing data, impoverishing the visualization. This finding led the authors to consider other solutions. Various approaches were explored, all reasonably limited considering the representation of temporal relations of precedence and simultaneity. As a result, a new visualization was created, EduVis, which proved to be adequate, consisting of the combination of a multi-layer display and a multi-matrix view.

Multi-Layer Visualization

This new approach, referred to as the multi-layer visualization, combines static and dynamic information in the same view. Each layer represents a semester of students' enrollment (not a semester of their curricula). For example, in Figure 3, three semesters are represented.

Figure 1. Minard's map (Napolean's campaign in Russian territory, 1812)

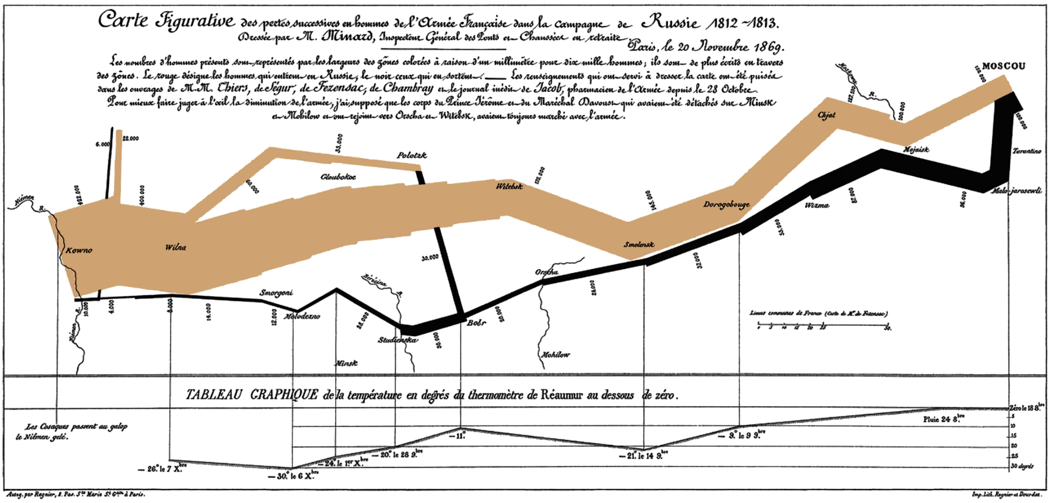
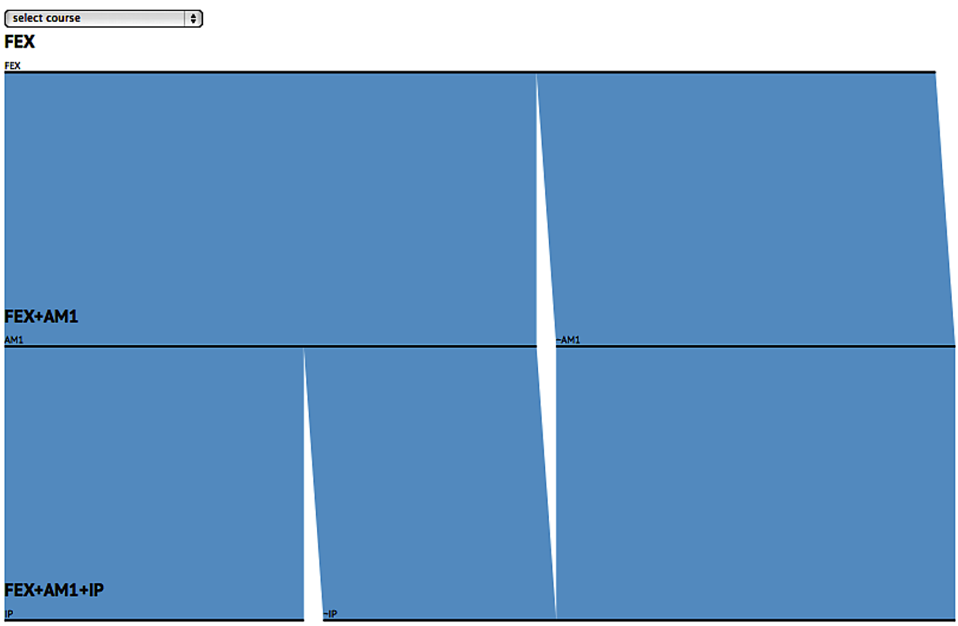
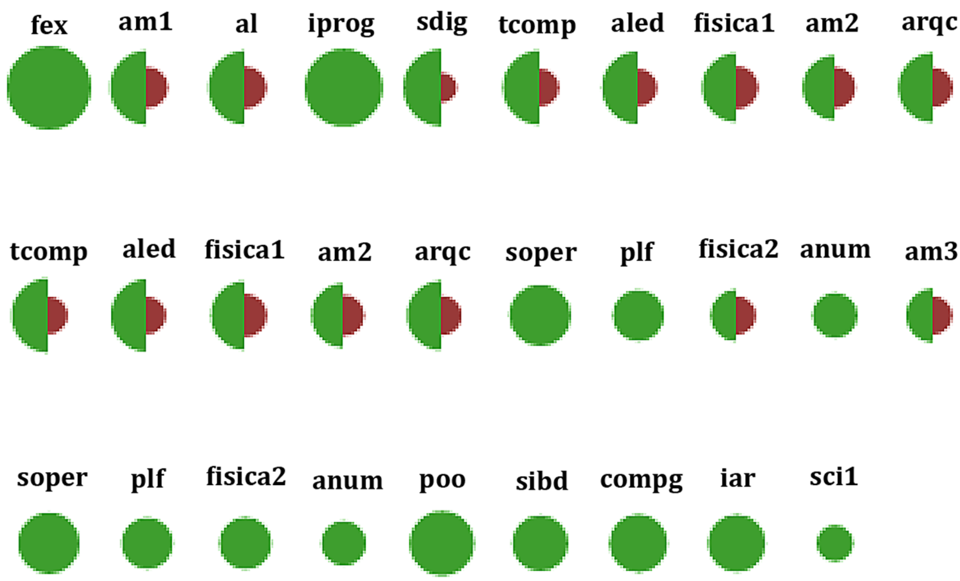


Figure 2. View inspired on Minard's map consisting of a Parallel Sets[1] approach



In this visualization, nodes representing each course correspond to circles, which are, if necessary, divided into semicircles, depicting the number of passing or failing students in a course. The radius of each (semi)circle is proportional to the number of students involved in the various patterns relating to that course. The computation of circles' radiuses takes into account, for standardization purposes, the maximum and minimum number of students who passed or failed a course. These limits are mapped to a range of pixels for display purposes. The minimum radius is fixed at four pixels (verified by experimentation to be the minimum value for a radius of a circle so that it is likely to be comfortably selected using a mouse). The maximum value is dynamically calculated and takes into account the

Figure 3. Multi-layer visualization



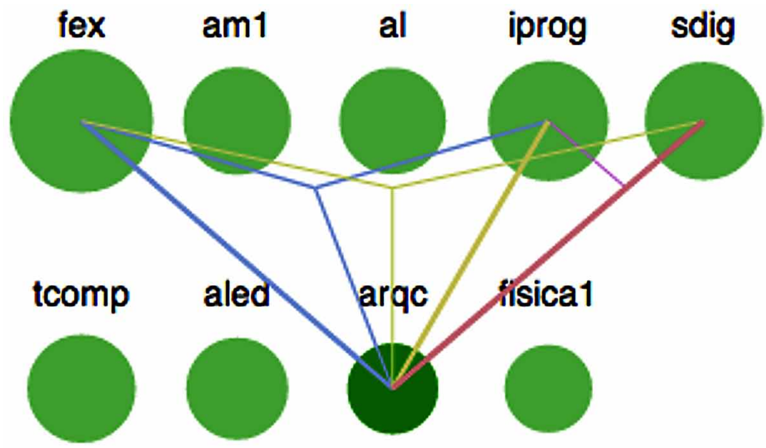
width of the display and the number of courses to represent. Thus, the radius of a particular course's circle is a linear mapping of the value of students who enrolled in that course to this range of visual values. For instance, in Figure 4 four courses are represented (*iar*, *am3*, *plf* and *pest*) in which *plf* has no support for failure (hence, only passing students are represented, consisting of a green circle), while *iar* displays lower rates of failure (smaller red semicircles) when compared to *am3* and *pest* (bigger red semicircles).

The multi-layer visualization allows the selection of multiple courses taken by students at a given time (horizontal), leading to restrictions which allow the immediate display of several possible routes followed by students along the curriculum (vertical). This selection is made by moving the mouse cursor over one course, resulting in the representation of the various patterns relating to it through arcs between various courses in each pattern. These connectors depart from the center of the selected (semi)circle and end at the (semi)circles corresponding to courses with which the currently selected one is interrelated, as illustrated in Figure 5.

Line thickness is proportional to the number of students involved. In Figure 5, the pattern *(sdig, fisica1)* corresponds to a higher number of students than *(fex, sdig, arqc)*. The computation of connectors' thickness is done through a normalization similar to the calculation of the circles' radiuses, that is, taking into account the maximum and minimum number of students present in the patterns to limit the thickness of the arc. The maximum thickness of the arcs prevents them to obscure the visualization as much as possible and is linearly mapped, taking into account the minimum value and the maximum pixel value of the diameter of the smallest circle.

Figure 4. Multi-layer visualization: representation of success / failure

Figure 5. Multi-layer visualization: representation of typical curricular student paths (patterns)



However, in some cases, the number of arches made the visualization unreadable, hiding the name of the subjects as shown in Figure 6 (left). Even though the authors tried placing the text labels over the visual connectors, these occluded part of the connectors, making it difficult to understand their organization. Hence, cubic Bezier curve connectors were implemented, as shown in Figure 6 (right). These curves, starting from the center of the selected course, are directed towards the center of the related courses. Control points are located in a vertical and/or horizontal distance from the selected course and each of the interrelated courses (Figure 7).

Furthermore, grouping mechanisms (bundling) for the arcs were created: multiple connectors that link the same set of courses are grouped into a single connector (Figure 8). The thickness of the resulting link corresponds to the standard value for the display of the sum of the students corresponding to each of the patterns. This normalization takes into account the range of values considered for the connectors, (minimum and maximum pixel values of the diameter of the smallest circle). User tests have demonstrated significant improvements in the performance of tasks with the use of Bezier curve connectors using bundling mechanisms [Gama 2014a].

Figure 6. Multi-layer visualization: linear connectors (left) and curve (right)

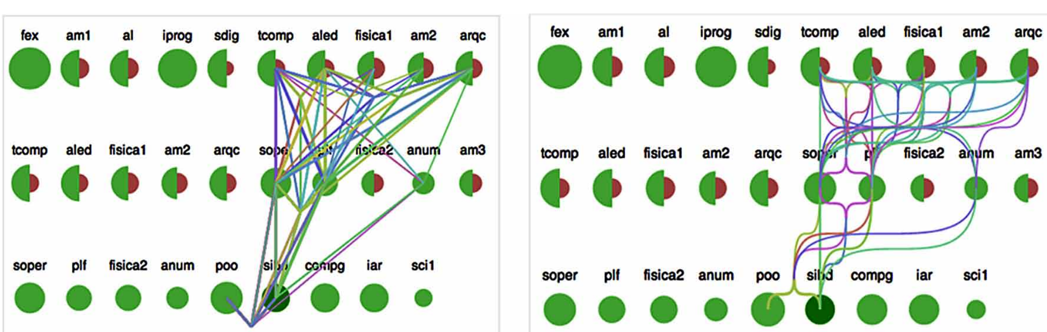


Figure 7. Multi-layer visualization: Connectors for aled: Bezier curves with control points in the vertical direction of the selected course circle's center and interrelated courses' centers. The vertical coordinates of control points correspond to half the distance between layers (semesters).

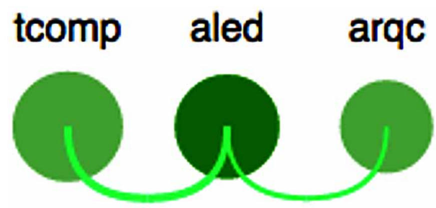
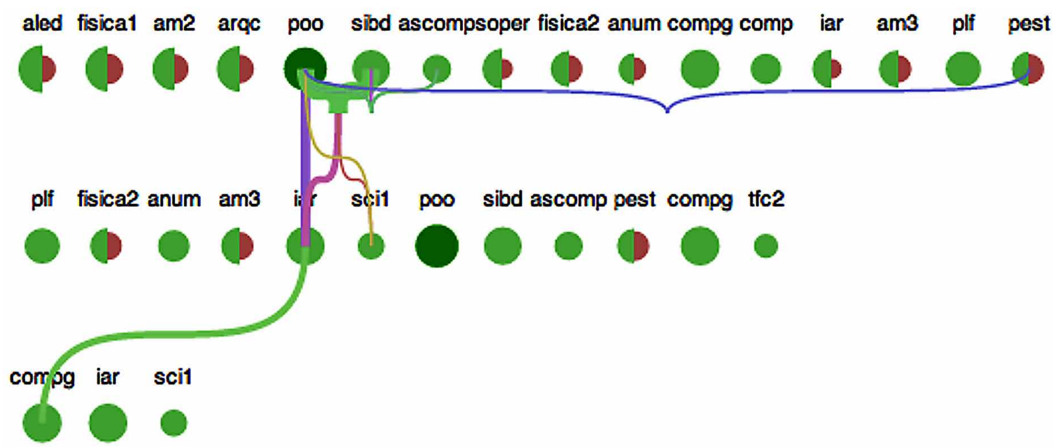


Figure 8. Multi-layer visualization: Grouping (bundling) of curved connectors



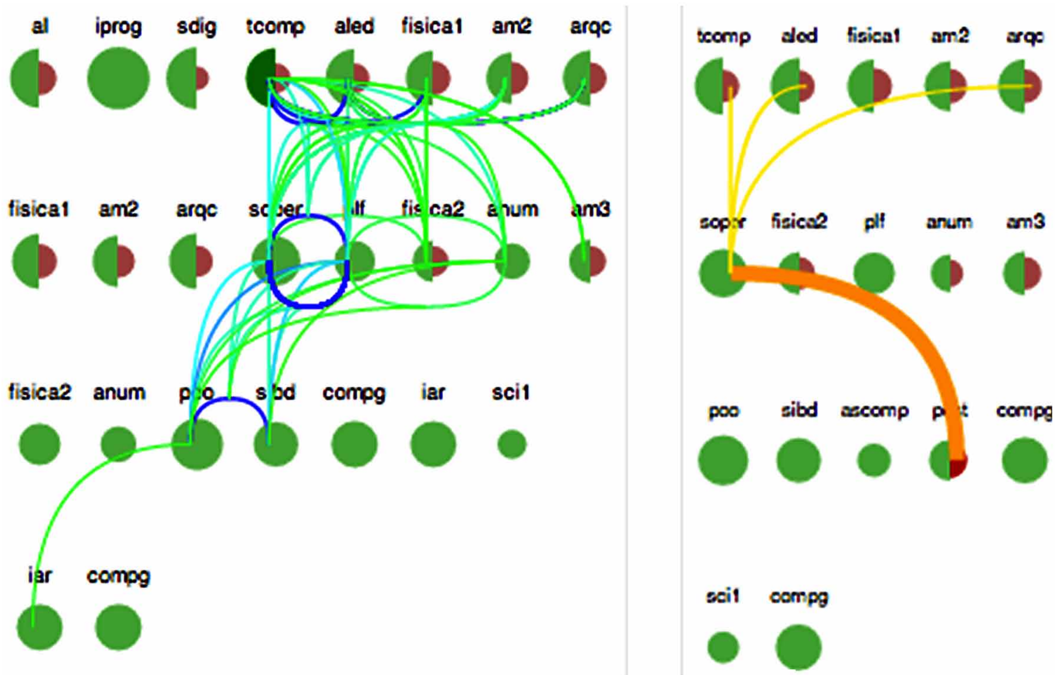
Color for Representing Information

Another problem encountered when considering the large number of visual connectors, is the fact that they contain a large quantity of information. Since they represent patterns associated with success and failure of the same course, such patterns must be distinguished.

Color was used to create this distinction: different colors were assigned to arcs related to passing and failing, and to offer redundant and complementary information on the number of students involved. Thus, the solution consisted of using color blending to enhance pattern representation. In particular, conventional Western color codes [Ware 2004] were used to represent success and failure: warm colors for failure and cool colors for approval.

In order to define a set of colors to use, a user study aimed at assessing the most natural color space for interpolation, which was the CIE-LCh, because of its perceptually uniform color characteristics [Gama 2014b][Gama 2014c]. In fact, certain pairs of colors produce better results when blended: (red, yellow), and (blue, green). The authors investigated the extent to which the human eye can distinguish the relative proportions of the original colors in color mixing [Gama 2014d]. The combination of the results of this set of studies allowed the correct association associate of color to success and failure. Even though if designing for color-blind people, different variations must be considered, the pairs (red, yellow), and (blue, green) were thus associated with warm (fail) and cool (pass) colors, respectively, and were thus mapped as shown in Figure 9. Thus, the greater the number of students corresponding to a pattern originating from a failed course, the closer the connector is to the color red (hot). Similarly, the greater the number of students corresponding to a pattern originating from a passed course, the closer the connector color will be to blue (cooler).

Figure 9. Color codes for success (left) and failure (right)



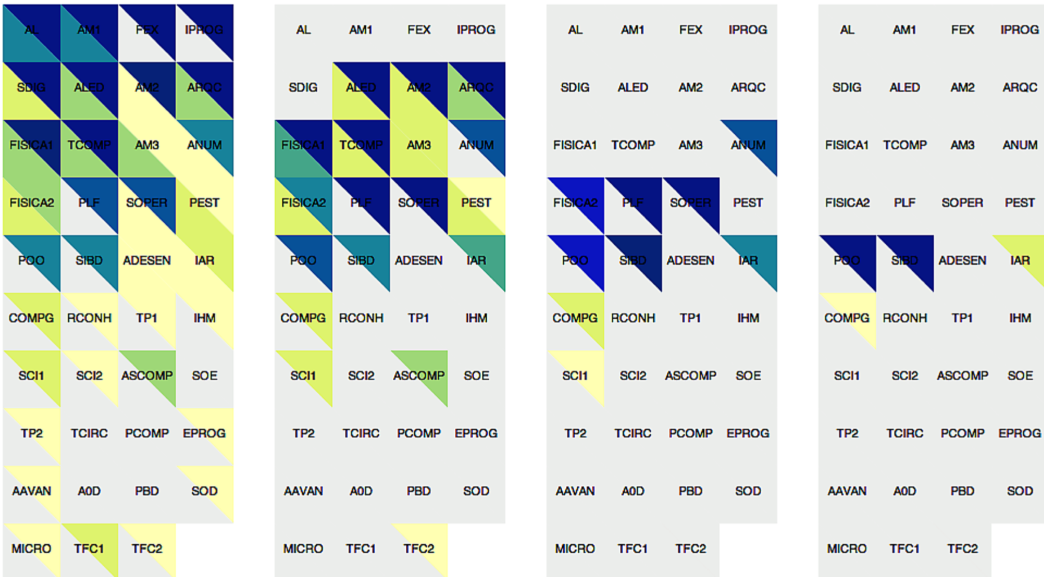
Multi-Matrix Visualization

Despite the aforementioned mechanisms allowing the easy exploration of static and dynamic aspects of patterns, it was not the most appropriate for a global visualization of the courses involved in a higher number of patterns (therefore more critical in the curriculum). To bridge this gap, the multi-matrix visualization was created (Figure 10).

In this representation, several matrices are presented side by side, corresponding to academic semesters. Each cell corresponds to a course, divided diagonally into two triangles: the upper triangle corresponds to course approval and the lower depicts course failure. The choice of the division of cells into triangles instead of rectangles was an attempt at maintaining visual continuity (secured by at least one contact point between two adjacent triangle cells), in contrast to the visual discontinuity that would be caused by rectangles, which would additionally lead to the visual artifact of vertical lines in the visualization. Furthermore, subdividing squares into triangles, using diagonal lines, avoided possible confusion associated with using vertical or horizontal subdivision as these are already used to differentiate course squares. The color intensity of each triangle is proportional to the number of patterns in which there is a reference to success (or failure) of the current course. So it is possible to have an immediate overview of the most important subjects (listed in a higher number of patterns). For instance, in the fist matrix on Figure 10, the topmost courses exist in a higher number of patterns (more intense colors, with lower brightness). The adopted color code in the multi-matrix visualization is consistent with that of the multi-layer representation.

This mechanism was integrated with the multi-layer visualization, resulting in a coordinated visualization mechanism, EduVis. The exploration of information related to a particular course takes place when the corresponding circle is selected. Since both views reflect all the curricular courses per semester, the selection may be made in any of the views by selecting a (semi) circle (multi-layer view) or a triangle (multi-matrix view). Thus, the multi-layer visualization shows the visual connectors

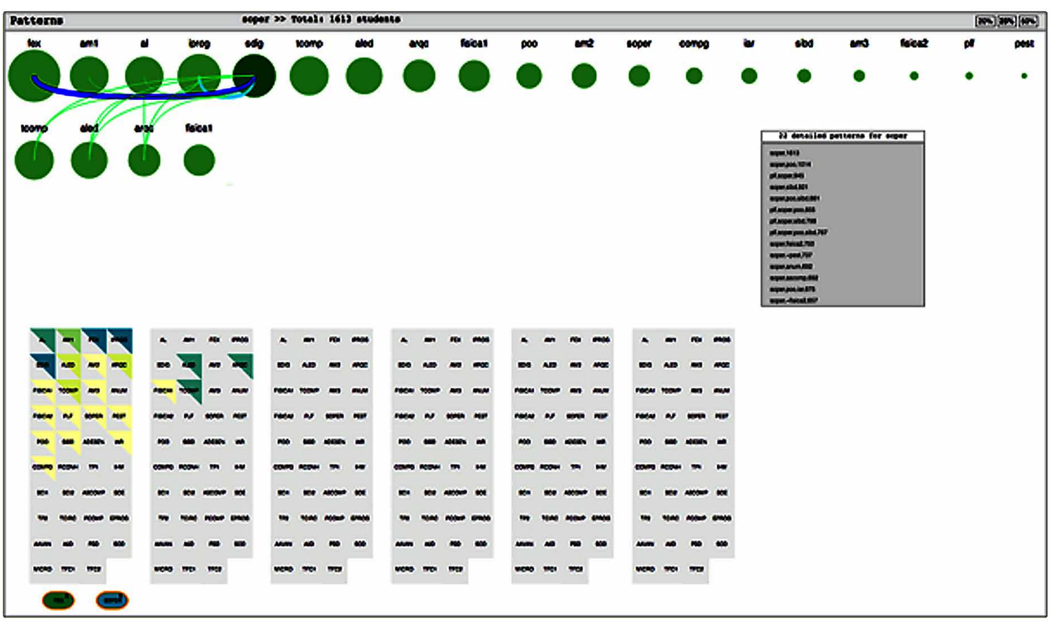
Figure 10. Multi-matrix visualization



corresponding to course interrelationships and the multi-matrix view shows these relationships by highlighting the courses involved with the currently selected, emphasizing dependencies (Figure 11).

When the mouse is pressed over a course and then leaves the corresponding circle, information previously shown continues to be depicted allowing simultaneous comparison of patterns. Subsequent pressure on multi-matrix's cells corresponding to other courses adds further constraints. This

Figure 11. EduVis



mechanism allows the creation of filters for information exploration that can be added and removed at any stage. Furthermore, the fact that restrictions can be removed in any order allows the exploration of various scenarios.

The matrix visualization after both *fex* and *~fisica1* have been selected as constraints in Figure 12. All other courses become gray, which provides context while keeping the focus on the selected courses. A first click on the triangle for *fex* highlighted all patterns in which passing *fex* was involved. One of those was *fisica1*. Clicking it leads to the situation in Figure 13.

Additionally, a specific reference may have two sets of constraints, each associated with a different color. Thus, cells belonging to two restrictions will be colored by blending the original colors.

Figure 12. Multi-matrix after selecting fex and fisica1

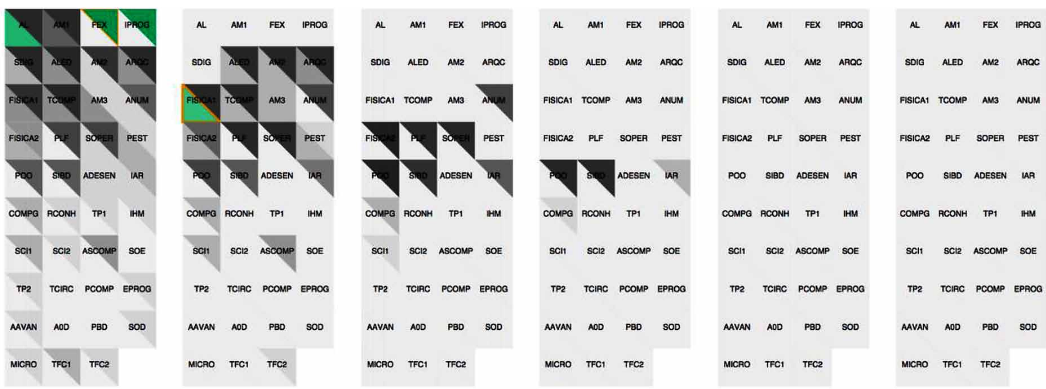
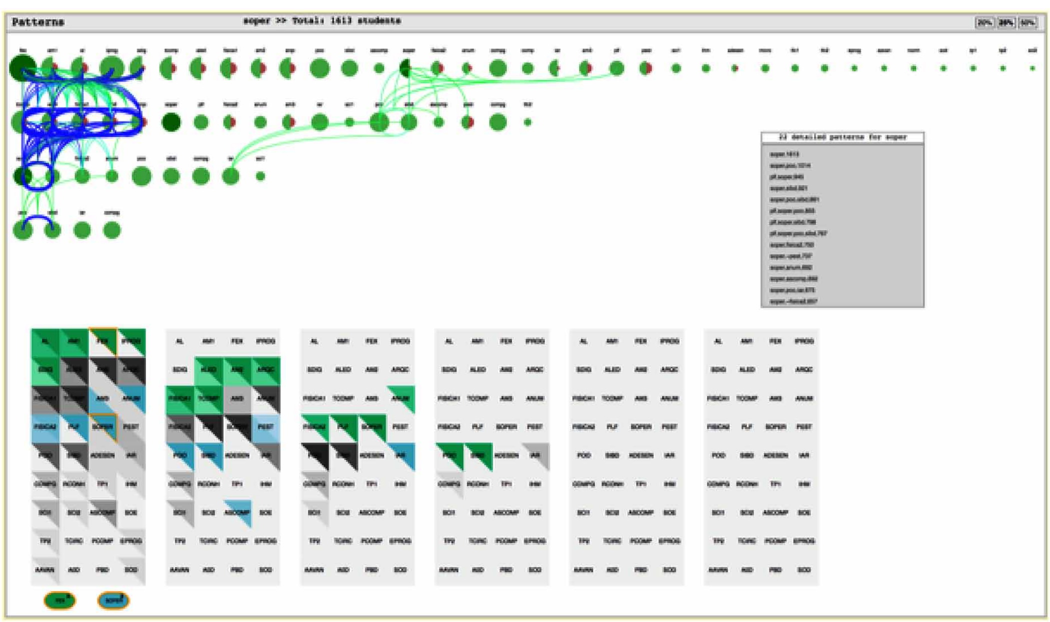


Figure 13. EduVis: Simultaneously comparing sets of patterns



EVALUATION

A user study was conducted to evaluate the created solution for visualizing educational patterns, aiming at ascertaining: (i) effectiveness and efficiency; (ii) usability in general and learning ability, in particular; (iii) degree of satisfaction in the performance of context-relevant tasks.

Representative Tasks

A set of tasks has been chosen to test the effectiveness and efficiency of EduVis in fulfilling the proposed objectives and thus showing that visualization may be successfully used to analyze information derived from EDM processes. So, the authors focused on the exploration of the visualization mechanisms in order to obtain all relevant information in this context.

Taking the represented information into account, a range of questions was designed to evaluate the visualization as a whole and its various mechanisms, as well as their integration and complementarity:

1. How many semesters are represented?
2. In general terms, what are the two courses with the most students?
3. What are the courses involved in the highest number of positive patterns in the 2nd semester?
4. What are the courses that are related to the *iar* course in the 4th semester?
5. Considering the students who enrolled in *am1* and *fex* in the 1st semester, what are the other courses in which they were successful in the 2nd semester?
6. What are the courses common to those who completed *sdig* in the 1st semester and failed in *aled* in the 2nd semester?
7. Considering the *po* course in the 4th semester, although it has more associated patterns, is it the course with more students in this semester?

The answer to questions 1, 2 and 3 should be inferred through visual analysis, allowing an immediate response. Question 1 may be answered by using any of the visualization mechanisms: the number of layers in the multi-layer view or the number of colored matrices in the multi-matrix display. Question 2 may be answered by observing, in the multi-layer view, the largest circles and the answer to question 3, although possible to provide using the multi-layer display view, requires further exploration, while it is immediate if the multi-matrix display is used (through the analysis of the second matrix, corresponding to the upper triangles with highest brightness).

The remaining questions require a higher degree of interaction. In question 4, to discover the courses related to *iar*, the user must move the cursor over the course circle (multi-layer display) or select its corresponding triangle (multi-matrix display). Questions 5 and 6 require the application of filters but in question 6, since there are overlapping courses, there is a color mix that the user should be able to distinguish. The visualization allows question 5 to be answered in two ways. The user may move the mouse over the selected courses and may immediately observe subsets of each course (these are highlighted with a color different from the selection), or, alternatively, (s)he may click the triangle of the course and, as a result, only the courses that are part of the corresponding subset are displayed in color.

Finally, the last question requires that the user combines information from both mechanisms in order to realize that a course, despite being involved in a high number of patterns (information from the multi-matrix mechanism), may not have a high number of students involved (information from the multi-layer mechanism).

Test Protocol

The aforementioned tasks were included in a questionnaire for conducting user tests, which were performed with 20 subjects selected through convenience sampling, with a maximum duration of

15 minutes, in a controlled laboratory environment with constant lighting conditions (as this could impact color perception, central to our visualization).

Before starting the evaluation, a comprehensive introduction was given to the participants, describing the context and objectives of the study. An introduction to the visualization and a small demonstration of its features followed, after which subjects were given 5 minutes to explore EduVis before performing the evaluation tasks. Participants were then given a questionnaire to answer the aforementioned set of questions, taking into account information gathered while interacting with EduVis.

Subjects were observed during evaluation and task performance time and errors were measured. Finally, participants were asked to answer an online satisfaction survey, divided into two parts: the first part corresponds to the System Usability Scale (SUS) (Sauro 2011) and the second part relates to a small set of questions that aim at evaluating, also using a Likert 5-point scale, the degree of difficulty felt by participants to perform each task. This last part of the questionnaire aims at revealing the degree of understanding of the following aspects:

1. Number of represented semesters (notion of the existence and location of a timeline).
2. Courses with larger number of approvals and disapprovals;
3. Courses involved in a higher number of patterns, that is, with more interrelationships with other courses (identification of key courses for students' academic success);
4. Comparison of different courses' patterns.

Results

Out of the 20 subjects, 15 (75%) were male, while the remaining 5 (25%) were female. Most (50%) lie in the range between 25-34 years of age, while 3 (15%) are between 18-24 years, 3 (15%) between 35-44 and 4 (20%) have 45 years of age or more. In terms of education, 16 (80%) either attend or have completed higher education, while 2 (10%) have completed primary education and 2 (10%), secondary education.

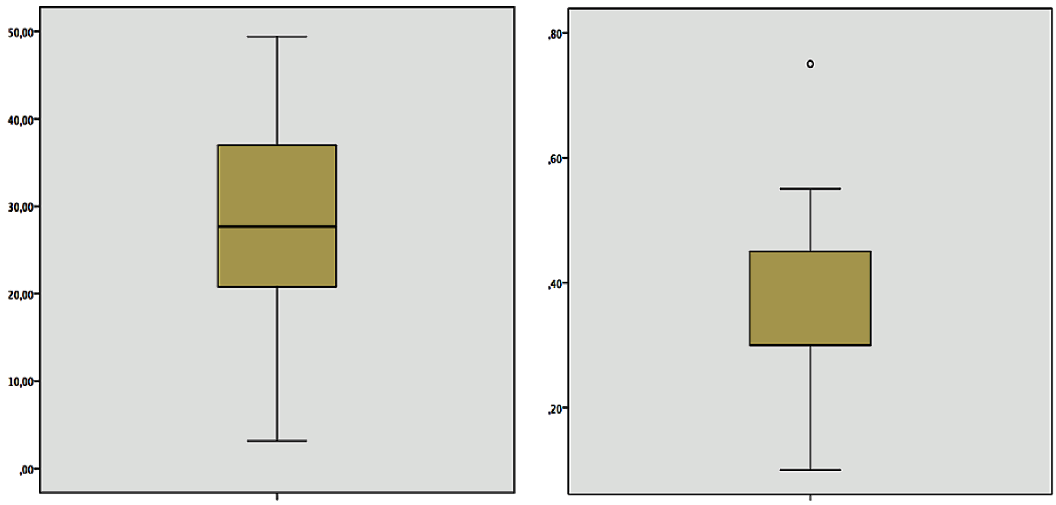
In terms of the solution's effectiveness and efficiency, time and average errors for task performance have been recorded and are summarized in Table 1 and Figure 14.

Given the time and the average number of errors on task performance, the measured values do not permit a clear distinction between the immediate tasks (1 to 3) and the exploration tasks (4 to 7), except for tasks 6 and 7. Such tasks seemed to take a considerably higher amount of time to complete, while task 6 also seemed to lead to a greater number of errors.

Table 1. Time and number of errors by task

Task	Time (s)		Nr. errors	
	Avg.	StDev.	Avg.	StDev.
1	3.20	1.25	0.35	0.48
2	17.60	12.56	0.10	0.30
3	28.30	15.80	0.55	0.59
4	23.95	13.11	0.30	0.46
5	27.70	9.71	0.30	0.46
6	49.45	28.25	0.75	0.62
7	45.70	25.28	0.30	0.71

Figure 14. Average distribution of time (left) and number of errors (right) for task performance



To further analyze these results, a statistical analysis was performed. The use of a Shapiro-Wilk test led to evidence against a normal distribution in most cases ($p < 0.05$), which suggested the use of a nonparametric test for further statistical analysis. Hence, a Wilcoxon signed-rank test was used to find significant differences between samples. Indeed, regarding time, subjects took significantly less time to perform task 1 than the others ($z_{1-2} = -3.72$, $z_{1-3} = -3.92$, $z_{1-4} = -3.92$, $z_{1-5} = -3.92$, $z_{1-6} = -3.92$, $z_{1-7} = -3.92$, $p < 0.05$), while they performed the second task significantly faster than tasks 3, 5, 6 and 7 ($z_{2-3} = -2.39$, $z_{2-5} = -2.54$, $-Z_{2-6} = -3.32$, $z_{2-7} = -3.25$, $p < 0.05$). Moreover, time taken to perform tasks 3, 4 and 5 was significantly less than to perform tasks 6 and 7 ($z_{3-6} = -2.31$, $z_{3-7} = -2.17$, $z_{4-6} = -3.45$, $z_{4-7} = -2.76$, $z_{5-6} = -3.36$, $z_{5-7} = -2.63$, $p < 0.05$). However, with regard to the number of errors, there are significant differences only between task 6 and those presenting lower error averages, task 2 ($W = 6$, $cv = 17$, $p < 0.05$), and task 4 ($W = 0$, $cv = 3$, $p < 0.05$), as well as between tasks 2 and 3 ($W = 5$, $cv = 8$, $p < 0.05$).

In an attempt to find a correlation between performance time and number of errors, Pearson coefficients were computed. A correlation was found, although relatively weak, in tasks 2 ($r = 0.45$, $p < 0.05$) and 7 ($r = 0.49$, $p < 0.05$), which makes it impossible to generalize a correlation between task performance and number of errors. Thus, the authors conclude that the temporal complexity does not lead to an increased number of errors when retrieving information using EduVis.

Regarding the satisfaction questionnaire, the score relating to the SUS according to corresponding calculation parameters (Sauro 2011), was of 79.47 points, indicating quite high results with regard to usability of the system. Using the same method to calculate the contextual satisfaction responses resulted in a score of 92.11, showing a very high degree of satisfaction in the performance of tasks.

The main objective of this work was to study a way to use visualization to make educational patterns evident while allowing further analysis and navigation. In particular, it was relevant to highlight success and failure in students' curricula and to provide mechanisms to compare and interrelate different courses. A list of tasks has been created to evaluate the created solution through user tests aiming at ascertaining the fulfillment of these objectives, in terms of effectiveness, efficiency, usability and satisfaction. In fact, test results validated the relevance of the created visualization, demonstrating the success of the solution in meeting the objectives: (i) effectiveness and efficiency; (ii) usability in general and learning ability, in particular; (iii) degree of satisfaction in the performance of context-relevant tasks.

CONCLUSION

A large amount of data emerges from the educational activities with the growing number of students in traditional and online education. An effective and thorough analysis of this information can help refine educational processes. In this context, data mining techniques have shown promise in finding relevant patterns in data. Alas, their application often leads to large sets of patterns, difficult to read, interpret and analyze. Only by overcoming these limitations will it be possible to study all that information as a coherent whole, discovering relevant features in the data and use them as a tool to support decisions and shape policies. Given this challenge, the authors created an interface where two different but linked visualization techniques complement each other. This allows the display and exploration of interdependencies of several college courses, resorting to side-by-side comparison, interactive exploration and filtering mechanisms. This system, EduVis, also takes advantage of colors and color blending techniques to emphasize relevant information, alleviating the visual clutter associated to the display of a large number of patterns with similar properties.

User tests have shown that our visualization allows the understanding of an overall view, as well as details about particular patterns and sets of patterns of special relevance for the problem domain. Users were able to understand the different semesters displayed, the sequential nature of patterns and to identify critical courses, with particularly high or low student success rates. The available filtering and comparison mechanisms were used and well understood. They also reported a good user experience and satisfaction when performing the tasks. These results showed that the goals that had been established for EduVis have been met. Possible uses of EduVis include the identification of critical points in curricula by degree coordinators striving for a better balance and the analysis by students themselves of possible troublesome points that may warrant additional study or care.

As future work, the authors intend to explore and extend the visualization techniques for patterns occurring within courses. In some contexts, courses have several evaluation moments (quizzes, homework, labs, etc.). The dependencies between these moments leading to an eventual outcome in terms of grade have also been the subject of study for educational datamining, again yielding sequential patterns. The authors expect to be able to apply their techniques to that domain. Furthermore, it would be interesting to establish prerequisites between courses to further explore patterns in a more meaningful way. Also, a further exploration of the use of color and color blending as well as more sophisticated edge bundling techniques will be explored to convey more information and increase understandability. Moreover, the authors plan to perform user tests and interviews with intended users, which will allow them to further evaluate their solution. Finally, these techniques will be applied to sequential patterns not related to the education domain, for further validation of the generality of the solution.

ACKNOWLEDGMENT

This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project Educare - PTDC/EIA-EIA/110058/2009.

REFERENCES

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of the IEEE International Conference on Data Engineering* (pp. 3–14). IEEE.

Antunes, C. (2008). Acquiring background knowledge for intelligent tutoring systems. *Proceedings of the International Conference on Educational Data Mining* (pp. 18–27). EDM.

Brooke, J. (1996). *SUS:* A quick and dirty usability scale. In *P.W. Jordan, B. Thomas, B.A. Weerdmeester, & A.L. McClelland (Eds.), Usability Evaluation in Industry*. London: Taylor and Francis.

Gama, S., & Gonçalves, D. (2014a). Visualizing Large Quantities of Educational Datamining Information. *Proceedings of the 18th International Conference on Information Visualization*. IEEE. doi:10.1109/IV.2014.65

Gama, S., & Gonçalves, D. (2014b). Guidelines for Using Color Blending In Data Visualization. *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. doi:10.1145/2598153.2600039

Gama, S., & Gonçalves, D. (2014c). Studying Color Blending Perception for Data Visualization. In *Proceedings of Eurovis, 2014*. Eurographics.

Gama, S., & Gonçalves, D. (2014d). Studying the perception of color components relative amounts in blended colors. *Proceedings of the 8th Nordic Conference on Human- Computer Interaction*. ACM. doi:10.1145/2639189.2670264

Gao, Y. (2010). A study on Mass Higher Education in the world Based on Comparative Perspectives. *Proceedings of the International Conference on Education and Management Technology (ICEMT)* (pp. 528-530). IEEE. doi:10.1109/ICEMT.2010.5657601

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A Tour Through the Visualization Zoo. In *Communications of the ACM, 5(6)*. ACM Press. doi:10.1145/1743546.1743567

Iliinsky, N., & Steele, J. (2011). *Designing Data Visualizations - Intentional Communication from Data to Display*. O'Reilly.

Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). MOOCs: So many learners, so much potential. *IEEE Intelligent Systems*, *28*(3), 70–77. doi:10.1109/MIS.2013.66

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

Mazza, R. & V. Dimitrova V. (2005). Generation of Graphical Representations of Student Tracking Data in Course Management Systems. *Proceedings of the 9th International Conference on Information Visualization* (pp. 253-258). IEEE. doi:10.1109/IV.2005.54

Mazza, R., & Botturi, L. (2007). Monitoring an Online Course with the GISMO Tool: A Case Study. *Journal of Interactive Learning Research*, *18*(2), 2007.

Romero, C., & Ventura, S. (2010). Educational Data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *40*(6), 2010. doi:10.1109/TSMCC.2010.2053532

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

Trimm, D., Rheingans, P., & desJardins, M. (2012). Visualizing student histories using clustering and composition. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2809–2818. doi:10.1109/TVCG.2012.288 PMID:26357190

Ware, C. (2004). *Information Visualization, Perception for Design* (2nd ed.). Morgan Kaufmann Publishers.

Wortman, D., & Rheingans, P. (2007). Visualizing trends in student performance across computer science courses. *Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education* (pp. 430–434). ACM Press. doi:10.1145/1227310.1227458

Xiaohuan, W., Guodong, Y., Huan, W., & Weu, H. (2013). Visual exploration for time seriesdata using multivariate analysis method. *Proceedings of the 8th International Conference on Computer Science Education (ICCSE)* (pp. 1189–1193). IEEE.

Xiaoya, G., Kan, L., & Ping, L. (2009). Visual analysis of college students' scores in English test. *Proceedings of the 4th International Conference on Computer Science Education (ICCSE)* (pp. 1816–1819). IEEE.

ENDNOTES

[1] Based in http://d3js.org/

Vilma Rodrigues Jordão holds a Master's Degree in Information Systems and Multimedia. Sandra Gama is a researcher at the Visualization and Multimodal Interfaces Group of INESC-ID and Professor of Computer Science at Instituto Superior Técnico (IST/UL), the School of Engineering of the University of Lisboa, Portugal. Her research interests fall within the scope of Information Visualization, Human Factors in Human-Computer Interaction and Multimodal Interfaces.

Daniel Gonçalves is a researcher at the Visualization and Multimodal Interfaces Group of INESC-ID and Professor of Computer Science at Instituto Superior Técnico (IST/UL), the School of Engineering of the University of Lisboa, Portugal. His research encompasses the areas of Information Visualization, Personal Information Management and Accessibility. He has published over 140 peer-reviewed papers publications in those areas, as well as being the co-author of a textbook on Human-Computer Interaction. He managed INESC-ID's participation in the national EDUCARE project, where information visualization techniques will be used in novel ways as a pedagogic aid, and EU AAL PAELife, striving to find new interaction techniques for the creation of a personal life assistant for the elderly. He was one of the organizers and Technical Program Chair of IFIP INTERACT 2011 (September 2011) and of the ACM International Conference on Intelligent User Interfaces (February 2012). He is a member of ACM and the Portuguese Computer Graphics Group (the national Eurographics chapter).