Use of reorderable matrices and heatmaps to support data analysis of students transcripts

Rafael Tavares Carvalho Barros, Thiago Gonçalves Mendes, Celmar Guimarães da Silva
School of Technology – University of Campinas
Limeira, SP, Brazil
r176257@dac.unicamp.br, t210905@dac.unicamp.br, celmar@ft.unicamp.br

***Abstract*—For a course coordinator, the analysis of several students' transcripts to identify the situation of subjects or students is often an old-fashioned process executed through a textual and numerical approach. This work is part of a larger project aimed at choosing appropriate visual representations to help course coordinators to analyze sets of students transcripts. In this work, we developed a system that allows the visualization of student transcripts through a *heatmap* of student grades per subject. The heatmap represent grades based on a user-defined color scale. To assist in the analysis, it is possible to reorder subjects and students using the *optimal leaf order* algorithm, or even to reorder according to the grades of a specific subject or student. In addition, some features have been developed to meet visual guidelines, such as overview, zoom, filter and details-on-demand.**

I. Introduction

The area of education produces a large amount of data that can be analyzed. Visualization resources can help to identify factors affecting the teaching and learning processes, enabling them to be improved. This may happen if the end users of these resources are able to understand the visualized data and information without much training [1] [2].

However, the existing process for analyzing data referring to a course or a set of students is often based on text and numbers, but not on interactive visualizations. Therefore, it is difficult or time consuming for a course coordinator to check if a course has a very different average score, to identify differences between different classes, or even to check if a student performed very differently in a given semester.

This work is part of a research that aims to help course coordinators in course analysis and in the identification of outliers through the visual representation of a set of student transcripts. Particularly in this work, we propose to visualize the transcripts with a reorderable heatmap, which displays grades of students in subjects according to a user-defined color scale. This may ease the perception of discrepancies, e.g. a subject in which the majority of the students performed poorly in a given year.

The work is organized as follows. Section II briefly presents a literature review on Educational Data Mining, heatmap and matrix reordering. Section III shows the methodology we adopted and some decisions for defining and constructing the proposed visualization. Section IV exemplifies the output of our tool for a synthetic dataset. Section V concludes our paper and briefly indicates future work.

II. Related work

This section presents two topics of interest: Educational Data Mining, and heatmap and reorderable matrix.

A. Educational Data Mining

The Educational Data Mining (EDM) literature references research areas such as data mining, data visualization, machine learning, and earliest works that apply Artificial Intelligence in Education [4]. In our literature review, we did not find studies that help to analyze student performance based on visualization of a set of transcripts. Some works use data from transcripts to forecast student problems, so they can try to avoid them [4] [5]. CourseViewer [14] visualizes a student transcript as a graph of subjects, subject prerequisites and grades, where subjects are nodes, prerequisites are edges connecting subjects, and node colors represent grades. However, this software shows only one transcript per screen.

B. Heatmap and reorderable matrix

A popular graphical visualization form is the heatmap, since it can represent large amounts of data in a small space for the analysis of possible patterns in the set [6]. It is a tabular graph that can represent at least two distinct types of datasets:

- A dataset with two independent categorical variables A and B and a dependent variable C, which may be ordinal or quantitative. Variables A and B are mapped to the columns and rows of the table. Variable C is mapped to the color of the cells.
- A multidimensional dataset. In this case, each variable is mapped to a distinct column of the heatmap, and each dataset tuple is mapped to a row.

A reorderable matrix [15] is a data structure that underlies a heatmap. Given that columns and rows of the heatmap may be permuted in both types of datasets without loss of data, finding a good permutation may help to unveil data similarities that could be hidden in the dataset. When interacting with the visual presentation, the user has the chance to detect patterns in the presentation and obtain more information about the data. This kind of pattern recognition is something that human vision is known to do remarkably well [3].

Matrix reordering can be done manually, where the user swaps rows and columns, or in an automated way, using algorithms that order the entire matrix or just the desired row or column. The principle of automatic visual reordering is

that rows and columns are rearranged according to their visual similarity.

Some examples of matrix reordering algorithms are: FVS (Feature Vector-based Sort), which reorders a data matrix aiming to reveal simplex and equi-correlation patterns [11]; Block Reordering, which is specialized at make evident Block pattern when it is hidden inside a reorderable matrix [12]; and Polar Sort, which produces high-quality results for uncovering Band and Circumplex patterns [13].

III. Visualization Tool

In order to develop our tool, we first defined its requirements. The tool is based on Shneiderman's Visual Information-Seeking Mantra [9]. This is a very useful starting point for developing a visualization tool, since it was developed based on analysis of human perceptual ability and studies of other visual design guidelines. Those guidelines can be summarized as: "overview first, zoom and filter, then details-on-demand".

With this mantra in mind, the system must enable: (a) filtering a subset of the classes; (b) providing an overview of the graph by zooming out and a focus on items of interest with zoom in; (c) details-on-demand as the mouse pointer is placed over some cell.

The system receives a transcript in a format based on CSV file (with header), as exemplified in Fig. 1. The first four rows have metadata. The first row has the field separator character; in CSV files, it can be semicolon, as in the example, or comma. The second row has the number of columns to be grouped to represent the student; in the example, the number 2 indicates that the row values of the first two columns (RA and CLASS) should be grouped and considered as the user identifier. The third row has the name of the column that represents the class, to allow the application of the class filter.

The fourth row presents the header of the values, and the remaining rows have the students' grades per subject. In the example, there are five subjects and the grades of two students in these subjects.

The graph generated from this file can be seen in Fig. 2, where each row represents a student, each column represents a subject, and the cell color varies from dark red (grade 0) to dark blue (grade 10) with a gray midpoint (grade 6). The users can change this color scale by changing these colors and by defining another intermediate value.

The tool chosen to generate the graph was D3.js [10], a JavaScript library for manipulating data-based documents. D3 helps to display data in the form of graphics using HTML (for page content), SVG (for vector graphics), CSS (for aesthetics) and JavaScript (for interaction) .

```
;
2
CLASS
RA;CLASS;SUBJECT1;SUBJECT2;SUBJECT3;SUBJECT4;SUBJECT5
000001;T2016_1;1;2;3;4;5
000002;T2016_2;6;7;8;9;10
```

Fig. 1. Example of an input file with two students and five subjects.

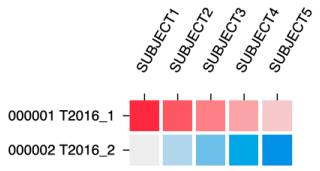

Fig. 2. Heatmap based on Fig. 1's file.

In order to generate the visualization, it was necessary to adapt the format of the input data to the format expected by D3.js. Thus, we generate a representation of the data in the form of an array of objects, each object containing the attributes *row*, *column* and *value*.

The developed system performs only automated rearrangements. We used a matrix reordering algorithm already implemented in the Reorder.js library [7], called OLO (*optimal_leaf_order*). The OLO method starts with a hierarchical clustering of rows (or columns); then it finds an order that is consistent with the dendrogram generated by the clustering and that minimizes the sum of distances between consecutive rows (or columns). We chose this method because it provides good results with only two parameters: the distance metric and the linkage type used for the hierarchical clustering.

In order to perform the heatmap reordering with the use of Reorder.js, a different data treatment than that used in D3.js was necessary. The tool generates two distance matrices (one for the rows and one for the columns), which are informed to the OLO method as parameters.

A. Example

Figs. 3 to 5 show visualizations of a synthetic dataset of students, subjects and grades.

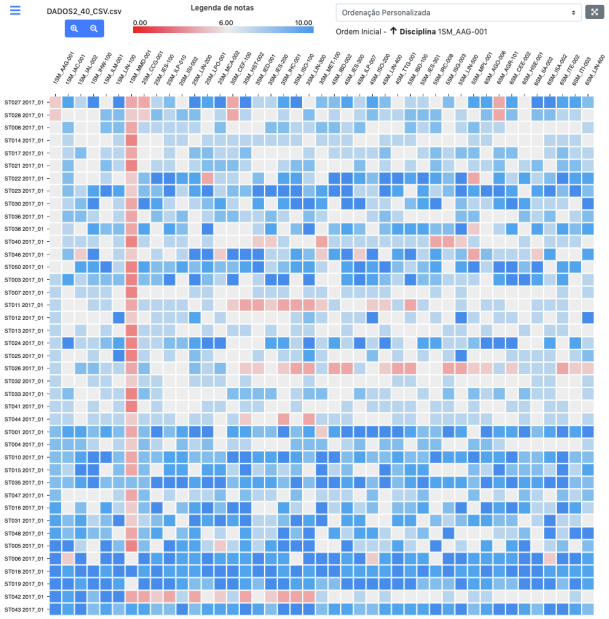

Fig. 3. Toolkit overview with heat map generated of 40 student transcripts, sorted alphabetically and by course subjects throughout the semester.

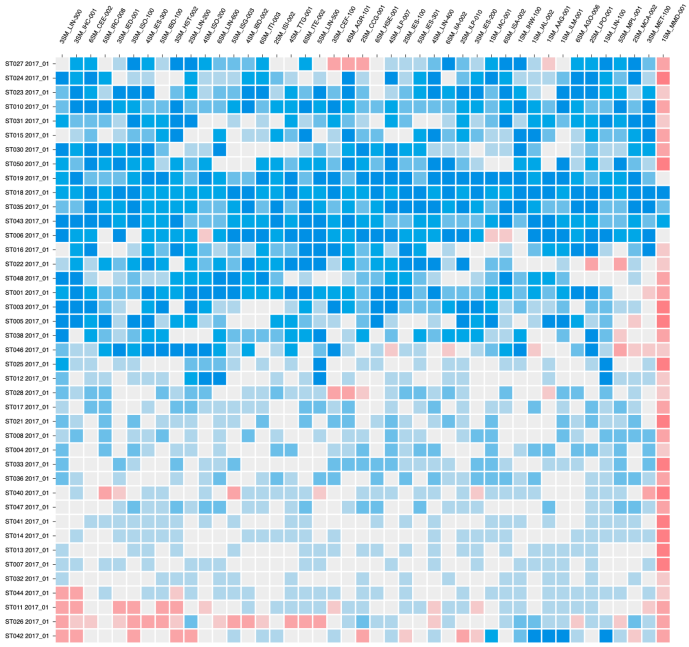

Fig. 4. Heatmap with the same data shown in Fig. 3, after beign reordered by OLO.

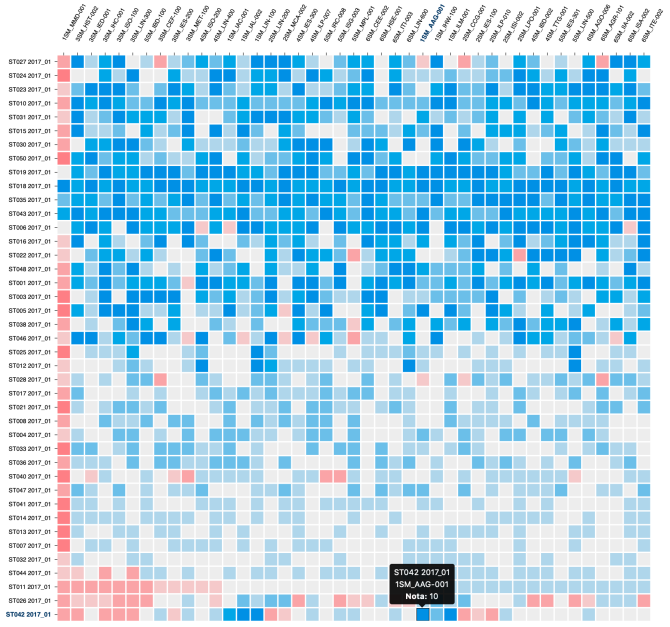

Fig. 5. Graphic sorted by ST011's grades (fourth row in bottom-up direction) . At the bottom, a detail window with the student ST042's grade in the subject 1SM_AAG_001.

The original input matrix (Fig. 3) is sorted by subject name and by student name. It shows some global patterns, such as a red column, that indicates a subject with median to low grades for almost all students. Coordinators may be interested in this pattern in order to understand which are the most hard subjects in his course in the student viewpoint. Other pattern present in this figure is a set of blue rows that refer to students with good grades in almost all subjects. Coordinators may select students with this performance for specific activities such as undergraduate researches.

Given that the subject names include year and semester, and that the subjects are sorted by time. This ordering may help users to observe bad situations in a period of a student (*e.g.*, the low grades in the third semester of the student ST011's transcript). This red-line pattern may alarm coordinators about possible personal problems of a student, such as psychological ones.

Fig. 4 has columns and rows ordered by OLO. It reveals distinct groups of students, according to their performance: top rows group students with the best grades in all subjects, bottom rows cluster students with the worst grades, and intermediate grades are in the central rows. This visualization may help coordinators to assess the overall performance of a course.

In addition, the system also enables a user to sort (in increasing or decreasing order) the heatmap rows according to the grades of a specific subject, and the columns according to the grades of a specific student. This ordering may help coordinators to detect students with correlated grades, and probably with the same possible behavior and presence or absence of problems. It may also aid coordinator in perceiving unknown correlations between subjects.

IV. Conclusion and Future Work

We implemented a system that performs the reading of students transcripts according to a specific format, and that generates a visualization of these data in the form of a reorderable heatmap. It aims to allow a better and faster visual perception on the transcripts, when compared to the textual analysis.

Based on the Visual Information-Seeking Mantra, the system enables class filtering, graphical overview and specific view with zoom manipulation, details-on-demand when hovering the mouse over a table cell representing the grade, and graph reorder. These features were developed aiming to improve the coordinator's experience, easing to observe data.

The use of the developed system is necessary to be able to affirm if the chosen graph, as well as the implemented features, helps coordinators to analyze sets of transcripts.

A discussion about the system is valid and encouraged, as well as its use in real cases to verify the actual utility of it and list enhancements to it, such as new functions or conversions in data entry for the file model that is interpreted, covering a greater number of universities or institutes that could use the developed system.

Some features that can be added to the system are: identification and grouping of subjects according to their similarities (e.g. mathematics, physics, sociology, programming); display information about the subject when selecting it; identify and inform the user about subjects where there were a number of above-normal failures; new reordering options.

Furthermore, the system is flexible enough to accommodate enhancements for more reorder algorithms.

V. Acknowledgements

We thank the grant #3221/19 from FAEPEX/PRP/University of Campinas.

References

[1] H. Qu, and Q. Chen, "Visual analytics for MOOC data". IEEE Computer Graphics and Applications, vol. 35, pp. 69-75, 2015.

[2] P. D. Ritsos, and J. C. Roberts, "Towards more Visual Analytics in Learning Analytics". Proceedings of EuroVis Workshop on Visual Analytics, pp. 61–65, 2014.

[3] H. Siirtola, and E. Mäkinen, "Constructing and reconstructing the reorderable matrix". Information Visualization, vol. 4, pp. 32-48, 2005.

[4] M. O. Hegazi, and M. Abugroon. "The State of the Art on Educational Data Mining in Higher Education". International Journal of Computer Trends and Technology, vol. 31, pp. 46-56, 2016.

[5] O. B. Coelho, and I. Silveira. Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review. Proceedings of the XXVIII Brazilian Symposium on Computers in Education (SBIE 2017), pp. 143-152, 2017.

[6] J. N. Weinstein, "A postgenomic visual icon". Science, vol. 319, pp. 1772–1773, 2008.

[7] C. Perin, P. Dragicevic, and J. D. Fekete, "Revisiting Bertin matrices: new interactions for crafting tabular visualizations". IEEE Transactions on Visualization and Computer Graphics, vol. 20, pp. 2082–2091, 2014.

[8] J.-D. Fekete. Reorder.js: A JavaScript Library to Reorder Tables and Networks. IEEE VIS 2015, Chicago, United States, 2015.

[9] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations". Proceedings 1996 IEEE Symposium on Visual Languages, pp. 336–343, 1996.

[10] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents". IEEE Transactions on Visualization & Computer Graphics 17(12), pp. 2301–2309, 2011.

[11] C. G. da Silva, B. F. Medina, M. R. da Silva, W. H. Kawakami, and M. M. N. Rocha, "A fast feature vector approach for revealing simplex and equi-correlation data patterns in reorderable matrices". Information Visualization, vol. 16, pp. 261–274, 2017.

[12] A. M. dos Santos, B. F. Medina, and C. G. da Silva, "Block Reordering: um algoritmo para evidenciar padrão Block em matrizes reordenáveis". Proceedings of the Workshop of Undergraduate Works (WUW) in the 29th Conference on Graphics, Patterns and Images (SIBGRAPI'16), 2016.

[13] C. G. da Silva, "Polar Sort: combining multidimensional scaling and polar coordinates for matrix reordering". Proceedings of the International Conferences Interfaces and Human Interaction 2019; Game and Entertainment Technologies 2019; and Computer Graphics, Visualization, Computer Vision and Image Processing 2019, pp. 239-246, 2019.

[14] F. B. Luiz. "Adequação da ferramenta CourseViewer para prover suporte à exibição de disciplinas extracurriculares e visualização por nota em disciplina". Undergraduate final work, School of Technology, University of Campinas, 2017.

[15] J. Bertin. Semiology of graphics: diagrams, networks, maps. University of Wisconsin Press, London, 1983.