Towards visual analysis techniques for monitoring students of distance education courses

Augusto Weiand, Isabel Harb Manssour
PUCRS, Faculdade de Informática, Programa de Pós-Graduação em Ciência da Computação
Porto Alegre, Brasil
augusto.weiand@acad.pucrs.br, isabel.manssour@pucrs.br

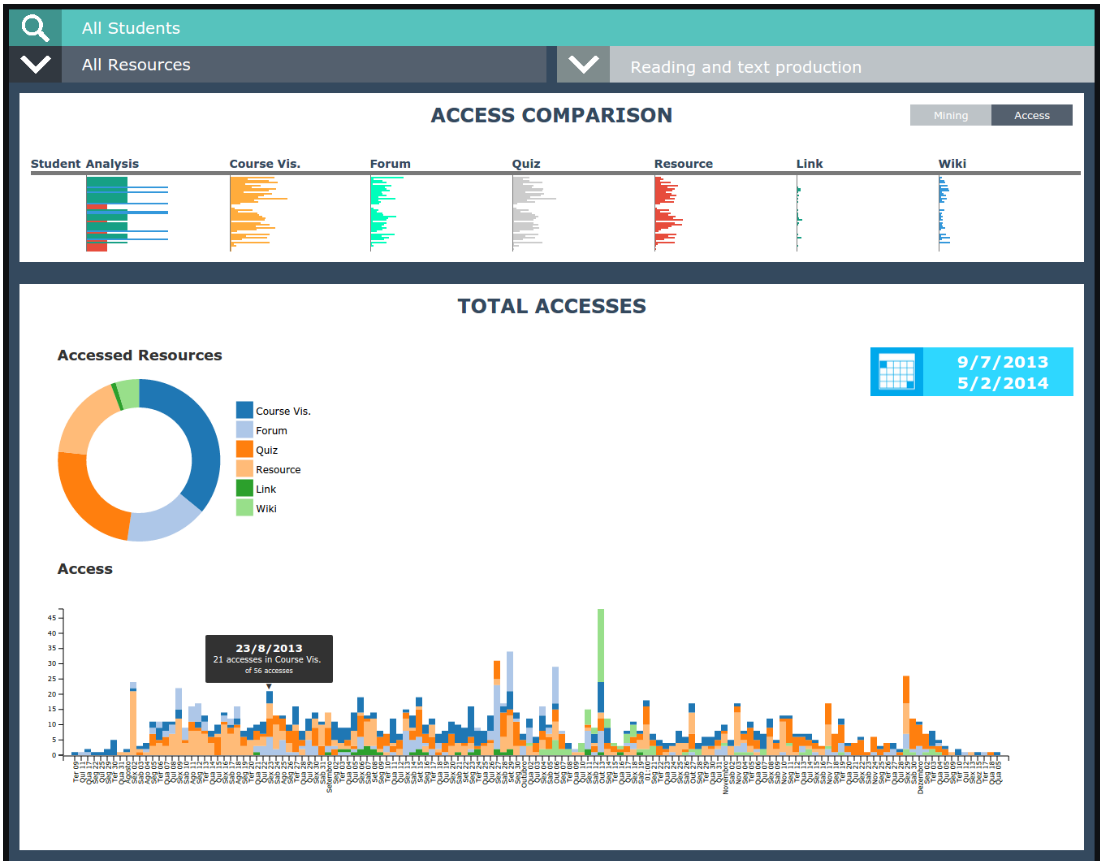

Fig. 1. Visualization example for data of 100 students: (top) used resources and the prediction result; (bottom) visualization for one student with mouse over interaction in the column graph.

***Abstract*—With technological advances, distance education has been much discussed in recent years, especially with the emergence of Massive Open Online Courses (MOOCs). The learning environments used in distance education courses usually generate a lot of data because of the large number of students and the various tasks involving their interactions. To facilitate the analysis of this data, and to allow for better monitoring of the students, we are doing a research to identify how interaction and visualization techniques integrated with data mining algorithms can assist teachers in predicting the students' performance in learning environments. The main goal of this work is to present the development of a visual analysis approach designed in this context that allows for assisting the gathering of data about the students' interactions, providing tools to investigate and predict the pass/fail rates in the courses.**

***Keywords*-visual analytics; distance education; data mining;**

I. Introduction

The popularization of the internet and advances in technology have brought along the exponential growth of distance courses. Due to the large number of students who join such courses, especially MOOCs, the amount of data generated by the learning environments they use has also increased. However, despite their growth, these courses face a major problem: student evasion [1].

Thus, in this context emerges the demand to allow for the analysis of the data generated by learning environments in order to keep up with the students' performance, since the amount of data hinders manual analysis [2]. Furthermore, as distance courses become more popular and more competitive, these analyses gain even more importance, for they enable organizations to develop work that is more suitable to each student's profile, as well as to keep abreast of their progress in the course.

One way to aid in the field of distance education consists of utilizing visual analysis techniques together with data mining algorithms [3]. Hence, with data from students' interactions with the learning environment, it is possible to identify the various student profiles, i.e. which of them show a tendency to evade, fail or pass. This type of prediction is made possible by the usage of mining algorithms to identify behavior patterns [1]. Similarly, the visualization of such information may be helpful in keeping to date with a large number of students and in analyzing the courses, providing for the discovery of the value added by the resources utilized and aiding in decision-making.

This work is aimed at describing an approach for keeping up with distance course students, integrating new forms of visual analysis and mining algorithms to the data generated by the employed learning environments. By means of a few visualization techniques, it allows teachers, tutors and administrators to be informed about their courses, delivering an overview of them, focusing on the students and the resources used in the environment. Our chief contribution is to provide a view of the resources made available in a course and their usage percentages among students, aside from allowing for an individual analysis of the students in a single interactive view.

The remainder of this work is organized in the following fashion: the next section provides a short description of some related work, as well as a few concepts related to this paper. Section III describes the proposed approach, and the last section presents final considerations and future work yet to be developed.

II. Background and Related Work

Distance Education: Distance Education is increasingly visible in the context of contemporary society as a more suitable model for the new educational demands, emerged from the personal and professional changes currently experienced in society [4]. In these circumstances, we can perceive an equal increase in the number of systems that seek to develop learning environments, so as to add value to the teaching-learning strategies [4], [5].

Data Mining: Data mining is defined [6], [7], [1] as a search for strategic information in data volumes much greater than a human can analyze, without, however, giving up the use of personal understanding of the available information, in order to make it useful for the organization. Moreover, data mining also employs prediction, originally from Artificial Intelligence, as a resource to aid in future data interpretation [8], [9], [10], [6]. After studying the literature [11], [12], [13], [14], [15], we have observed that some of the most popular algorithms for data mining in learning environments are decision trees, clustering, classification, k-means and naïve Bayes.

Visual Analytics and Distance Education: With the growing number of data currently generated, visual analysis techniques associated with data mining are being more and more used to aid in the pursuit of insights about a data set [16], [17]. Despite that, the study of related work endorses the perception that not always are these techniques combined in the field of distance education. Apparently, those works present relevant data mining algorithms, but the visualizations, when existing, seem to merely present the results in tables, without using intuitive and interactive graphical elements.

For instance, the work of Zhong-mei et al. [11] presented a combination of techniques for mining educational data to explore the results of students' learning. From that, they developed a learning result prediction model but did not make visualization available. Similarly, Cocea and Weibelzahl [14] utilized decision tree algorithms for the construction of a student motivation prediction system in distance courses without a graphical approach.

Johnson and Barnes [18], [19], on the other hand, proposed a data visualization tool that provided a graphical representation of a decision tree. The goal was to produce insights on the way students solve procedural domain problems. Some simple graphs were also supplied in the work of Romero et al. [13] to the end of presenting data about students registered on a learning environment such as grades, questionnaires and forums.

Mazza and Dimitrova [20] developed a tool called *CourseVis*, initially an extension of the Blackboard learning environment, but with a generic structure that would make it usable in other environments. Its design was conceived so as to allow instructors to check information on social, cognitive and behavioral aspects of their students, using line, bar and dispersion graphs. Through the environment's logs, information such as the view count for a forum topic is extracted, to then be used to present a graph of students who started the topics, for example. An analysis of this tool concluded that its graphical representations aided instructors in obtaining information on students more rapidly and with greater precision. Thus, the authors later developed a moodle plugin called GISMO, using the same visualization principles [21].

In the same way that the works mentioned presented approaches for mining and visualizing learning environment data, such tools cannot easily be found for MOOC environments. Most environments used for these courses are developed by universities themselves and, through research carried out to date, no environment was found that provided access to its administration area or to its data. Besides, no work was found in the literature that presented the use of an approach similar to the one discussed in this paper.

On the other hand, the usage of other web environment analysis systems is notorious, as is, for instance, *Google Analytics*[1]. This system provides tools to accomplish various

[1] https://www.google.com/analytics/

evaluations of the environment where it is installed, such as the path taken by students through the pages of a website, the number of views and the average duration of their stay. However, as discussed in Dragos' article [22], it has limitations concerning its use on educational systems. Since its goal is aimed towards e-commerce, as showed in the article, many discrepancies han be found when software such as *Google Analytics* and an appropriate analysis approach are used to assess the same educational dataset.

III. Visual Analysis Proposed

In order to develop the new approach for keeping up with distance course students, data mining techniques were put into practice to identify three student profiles in the courses to be analyzed: tendencies to pass, fail or evade. To aid in this analysis, two main views were tailored: one for the mined and categorized data, and another containing the interactions performed by the students. These views, as well as the algorithms and the data utilized, are described next.

A. Data Input and Algorithms

Different distance education courses employ different learning environments. In order for the proposed approach to not be restricted to any specific environment, and after carrying out a study on the data available in these environments, we have developed a set of default tables for internal use. As a result, data from various environments can be converted into this structure to be imported. Some of the data required to be included in these tables are the learning environment access records and data concerning the courses, such as usage of available resources, number of views and comments in forums, etc. Optional data include final (which can be used for training) and partial grades.

Research carried out in the literature [23], [20], [21], [24], [3], [25], [13] has demonstrated that moodle is used in several universities. Hence, initially, we developed a mechanism for importing moodle's database and we are running tests with this environment's data that was granted by Faculdade Cenecista de Osório (FACOS). This data is related to numerous distance education courses offered in the last three years by the institution. These courses were created with diverse interaction resources, including forums, wikis, quizzes and external URLs, and each of them contains data on 100 to 600 students.

For data mining, three algorithms were used: k-means, naïve Bayes and decision trees. The first two presented unsatisfactory results, and the latter, due to its characteristics, has facilitated validation and had better accuracy, being, therefore, the one selected. Table I presents the information regarding the performance of each algorithm.

Its usage requires previous training, which allows it to perform predictions and classifications. To that end, an annotated data set was produced, containing all the records on students' interactions and their classifications. After training, for its usage, the table data with access records is converted into the data format expected by the mining algorithm and by the developed visualization approach.

TABLE I
PERFORMANCE COMPARISON OF THE TESTED ALGORITHMS

	Correct %	Correct	Incorrect %	Incorrect
Decision Tree	76,00%	23	23,00%	7
K-means	30,00%	9	70,00%	21
Naïve Bayes	23,00%	7	76,00%	23

The methodology used for testing was as follows: after selecting a course from the dataset, 50 students were randomly removed, so as to not take part in the algorithm's training. After training, 30 of the 50 students were chosen, also randomly, to integrate the accuracy validation step, comparing the student's final grade with the classification found. If the student had a grade greater than 7, they passed; if between 5 and 7, they failed; if lower than 5, their access records would also be considered. If the number of accesses were 30% below the other student's average, they were deemed to have evaded the course.

B. Visualizations

After the data is input and mined, the processed data is then passed on to the visualization module, which produces the output illustrated in Fig. 1. In this figure, two distinct views can be noticed: one with the data mining algorithm's result (top of Fig. 1); and another with information about students' visits to the various environment resources (bottom of Fig. 1).

In order to provide for the displaying of information on several students without occupying too much space and without the need for scrolling, a view was implemented based on the table lens technique [26]. This view allows for a rapid analysis and comparisons across all students and resources without covering too much of the screen. Each line represents information about a single student and each column represents a resource of the course in question. By clicking on a line, it expands to reveal more details about the selected student, as shown in Fig. 2.

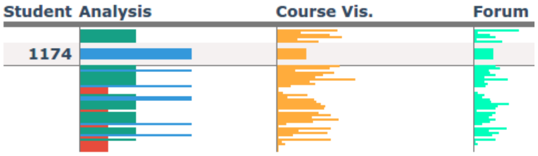

Fig. 2. Table lens with a line selected.

Despite that, it is known that, even using the table lens technique, a course with too many students may require page scrolling for the complete visualization in a monitor. However, since there is variation in monitor size and resolution, the use of larger monitors with better resolution allows for the visualization of a larger number of students without the need for scrolling. The results obtained in the present work demonstrate that a set of 400 students can be visualized without scrolling on a 14-inch display with a resolution of 1366x768 pixels.

The colors in the *Student Analyses* column (Fig. 1) represent the profile identified for the student: red for a student with tendency to evade, green for a student with tendency to fail and blue for a student with tendency to pass. The student profile identification is the result of the mining algorithm, based on the student's visits to the different resources and, possibly, the partial grades available in the course. The remaining columns have a single color and represent, in this example, the number of times each student visited each resource.

The visits view (bottom of Fig. 1) presents all the visits made by students in the course through time, grouped by resource visited. The column graph can handle time interval selections, which causes a recalculation of the proportions shown in the donut chart.

C. Interactions

Due to being an approach that seeks to provide a simpler and more intuitive view for the user, we developed several forms of interaction that enable broad data manipulation. The basic interaction resources are made available via the menu shown in Fig. 3. If a student's name is typed in the *All Students* field, a search is performed and the student's information is automatically highlighted, as illustrated by Fig. 2. This action also affects the column and donut graphs, so as to represent only the data concerning the selected student. Likewise, leaving the field empty causes the graphs to return to their initial state.

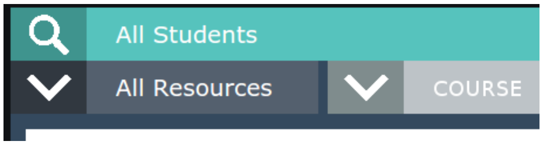

Fig. 3. Selection of basic interaction resources.

In this menu (Fig. 3), it is also possible to select one or more resources to be analyzed in various ways, or to let all of them be visible. Additionally, it is also possible to change courses.

Concerning the columns in the view at the top of Fig. 1, it is possible to click on one of them to effect the reordering considering the data in the column clicked, from highest to lowest and vice-versa. Furthermore, changing column placement is also possible.

Another implemented form of interaction allows the user to use the mouse to select a given period in the column graph, as can be seen in Fig. 4. This way, the donut chart will represent the total resource visit counts within the selected period. Moreover, when the cursor is over a bar, more detailed information about the selected bar is presented, including the total number of visits in a day and the number of visits regarding the resource. This occurs due to the graph being of the stacked kind. This way, when users place the cursor over another part of the bar, defined by another color, they will see the visits pertaining to the resource it represents.

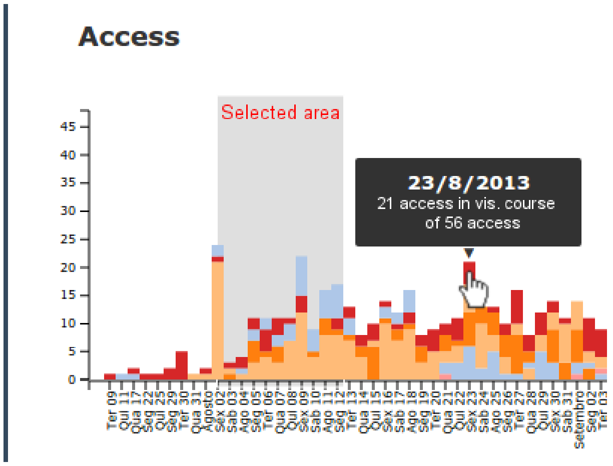

Fig. 4. Selected area and mouse over information in the stacked column graph.

IV. Conclusions and Future Work

An approach that integrates new forms of visual analysis and mining algorithms in educational environments was developed and presented in this work. Its goal is to provide an overview of distance education courses, aiding in reducing evasion and/or fail rates and in understanding the students' profiles.

At the moment, we are working on analyzing the produced results to improve the classifications performed. As future work for this approach, we intend to expand it for usage with data from various learning environments and to test it with potential users to validate the views and the interactions supplied.

Acknowledgment

This paper was achieved in cooperation with Hewlett-Packard Brasil Ltda. Using incentives of the Brazilian Informatics Law (Law n°. 8.248 of 1991). We would like to also thank Faculdade Cenecista Osório for the database provided for this research.

References

[1] A. J. C. Kampff, V. H. Ferreira, E. Reategui, and J. V. Lima, "Identifying evasion and poor performance profiles for the generation of alerts in a distance learning context," *American Latin Journal of Educational Technology*, vol. 13, no. 2, pp. 61–76, 2014.

[2] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, *Visual analytics: Scope and challenges*, 1st ed., ser. Lecture Notes in Computer Science, S. Simoff, M. Böhlen, and A. Mazeika, Eds. Springer Berlin Heidelberg, 2008, vol. 4404.

[3] C. B. Baruque, M. A. Amaral, A. Barcellos, J. C. da Silva Freitas, and C. J. Longo, "Analysing users access logs in moodle to improve e learning," in *Proceedings of the 2007 Euro American conference on Telematics and information systems*, ser. EATIS '07, no. Article 72, ACM. New York, NY, USA: Euro American conference on Telematics and information systems, 2007, pp. 72:1–72:4.

[4] P. L. Rogers, G. A. Berg, J. V. Boettecher, C. Howard, L. Justice, and K. Schenk, *Encyclopedia of distance learning*, 2nd ed. Information Science Reference, January 13 2009, vol. 4.

[5] A. T. C. Pereira, V. Schmitt, and M. Dias, "Virtual learning environments," *Virtual learning environments in different contexts*, vol. 1, p. 5, 2007.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed., S. A. Edition, Ed. Morgan kaufmann, July 2011, vol. 1.

[7] S. M. Weiss and N. Indurkhya, *Predictive Data Mining: A Practical Guide*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, vol. 129, no. 1.

[8] P. Adriaans and D. Zantinge, "Data mining," *England: Addison Wesley*, vol. 1, p. 1, 1996.

[9] M. J. Berry and G. S. Linoff, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

[10] R. Elmasri and S. B. Navathe, *Fundamentals of database systems*, 6th ed. USA: Addison-Wesley Publishing Company, 2014.

[11] S. Zhong-mei, Q. Qiong-fei, and F. Lu-qi, "Educational data mining and analyzing of student learning outcome from the perspective of learning experience," *Proceedings of the 7th International Conference on Educational Data Mining*, pp. 359–360, 2014.

[12] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering The Information Age-Solving Problems with Visual Analytics*, 1st ed., Goslar, Ed. Eurographics Association, 2011, vol. 7, no. 0, proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (FET 11).

[13] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use moodle courses," *Computer Applications in Engineering Education*, vol. 21, no. 1, pp. 135–146, 2013.

[14] M. Cocea and S. Weibelzahl, "Can log files analysis estimate learners level of motivation?" *14th Workshop on Adaptivity and User Modeling in Interactive Systems*, pp. 32–35, 2011.

[15] R. Pedraza-Perez, C. Romero, and S. Ventura, "A java desktop tool for mining moodle data," in *Proceedings of the 3rd Conference on Educational Data Mining*, ERIC. Educational Data Mining, 2011, pp. 319–320.

[16] J. J. Thomas and K. A. Cook, *Illuminating the path: The research and development agenda for visual analytics*, N. Visualization and A. Center, Eds. IEEE Computer Society Press, 2005.

[17] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, A. Press, Ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.

[18] M. Johnson, M. Eagle, L. Joseph, and T. Barnes, "The edm vis tool," in *Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, July 6-8, 2011*, Eindhoven, The Netherlands: International EDM Society. Educational Data Mining, 2011, pp. 349–350.

[19] M. Johnson and T. Barnes, "Visualizing educational data from logic tutors," in *Intelligent Tutoring Systems*, J. M. Vincent Aleven, Judy Kay, Ed., vol. 6095, Springer. Springer Berlin Heidelberg, 2010, pp. 233–235.

[20] R. Mazza and V. Dimitrova, "Coursevis: A graphical student monitoring tool for supporting instructors in web-based distance courses," *International Journal of Human-Computer Studies*, vol. 65, no. 2, pp. 125–139, feb 2007.

[21] R. Mazza and C. Milani, "Gismo: a graphical interactive student monitoring tool for course management systems," *International Conference on Technology Enhanced Learning, Milan*, pp. 18–19, Nov 2004.

[22] S. M. Dragos, "Why google analytics cannot be used for educational web content," in *Next Generation Web Services Practices (NWeSP), 2011 7th International Conference on*. IEEE, Oct 2011, pp. 113–118.

[23] S. Graf, C. Ives, N. Rahman, and A. Ferri, "Aat: a tool for accessing and analysing students' behaviour data in learning systems," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, ser. LAK '11, ACM. New York, NY, USA: ACM, 2011, pp. 174–179.

[24] A. Bovo, S. Sanchez, O. Héguy, and Y. Duthen, "Demonstration of a moodle student monitoring web application," *The 6th International Conference on Educational Data Mining*, pp. 390–391, 2013.

[25] M. I. Lopez, J. Luna, C. Romero, and S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums." *5th International conference on educational data mining*, pp. 148–151, 2012.

[26] R. Rao and S. K. Card, "The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '94. New York, NY, USA: ACM, 1994, pp. 318–322.