



Text mining of academic blogs for analysis of intellectual capital generated at the Faculty of Technology

Fabrício Talarico*, Antônio Carlos Zambon.

Abstract

The goal of this work was to develop the process to collect and analyze data from blogs used as a tool for students collaboration in active learning disciplines. In order to solve operational issues of managing the volume of assignments to grade and text quality monitoring, problems that blogs usually present, an educational data mining tool, based on natural language processing, has been developed. This is the result of a research developed since 2017 at UNICAMP- Brazil, which made it possible to collect sufficient information for the development of a system and preliminary evaluation of the results. The first results obtained were of noticeable improvement in the information processing capacity and the creation of a new process to analyze and monitor the evolution of theoretical and practical quality of the students learning in a discipline.

Key words:

natural language processing, educational data mining, intellectual capital

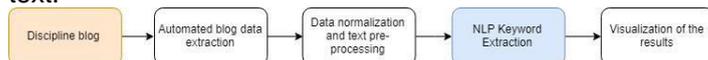
Introduction

Educational data mining (EDM) is the application of computerized methods to detect patterns in large collections of educational data that would otherwise be difficult or impossible to analyze (Baker, 2012). EDM analyses data generated by any type of information system supporting learning or education in academic and vocational education institutions. In this work the goal is to build an algorithm that recursively collect data building a database of students' interactions on internal blogs of the UNICAMP Faculty of Technology and uses natural language processing techniques to extract keywords from students texts, aiming to provide information on the evolution of theoretical and practical quality that constitutes the students of the disciplines. It is also worth mentioning that an article of this research was accepted to ECKM 2019 (<https://www.academic-conferences.org/conferences/eckm/>) in which all the analysis process that this work proposes is described.

Results and Discussion

The process created can be summarized in four main phases: automated data extraction, data normalization and text pre-processing, model building (NLP key word extraction) and visualization of results (Image 1).

Image 1. Process of extracting and analyzing students text.



To achieve this process was made a bibliographic study of EDM and learning analytics processes, approaches and applications. That study evolved until was clear that a NLP algorithm to extract the keywords of the student's texts would enable a quick and concise analysis of the students understanding of the disciplines topics. A TD-IDF machine-learning model was used to perform the keyword extraction. The result of that extraction was processed and integrated with a data visualization tool so the teacher of the discipline can analyze and monitor the interactions of the blog. The operationalization of a teaching-learning environment based on active

methodologies and collaborative learning can only be considered in this way, if it allows the free interaction of students to build their own knowledge. The Blog is highlighted in this type of initiative, as it encourages interaction. However, students need, in addition to unrestricted interaction, elements that can tutor them in the conceptions that are being formed over a given time. The teacher must have the means to, in addition to identifying deviations, also notify them in a timely manner, so that the whole learning effort is not dispersed or structured on the wrong concepts. Finally, unanswered questions or responses that are out of date discourage them, throwing away all efforts towards collaboration. The blog, supported by an NLP instrument, broadened the analytical opportunity of the teacher and enabled the achievement of previously stipulated pedagogical goals. The cumulative need to interact in the classroom and asynchronously in the blog is no longer a restriction factor in the application of the process, especially the NLP tool, which monitors the responses in the blog in real time.

Conclusions

This work presented a solution to a problem experienced in the academic environment, which refers to lack of information on the evolution of theoretical and practical quality of the students in a university course. To mitigate these problems, a tool was developed for analyzing and monitoring student discourse in blogs. The new dynamics of text analysis gave the teacher more agility, allowing them to broaden the scope. The teacher increased his capacity of management of the individual access and identification of the students with better argumentation.

1 Baker, K. (2012). Optimization modeling with spreadsheets. Retrieved from <https://bit.ly/2Jztrpl>

2 Facer, K. (2011) Learning futures: Education, technology and social change, Learning futures: Education, technology and social change. doi: 10.4324/9780203817308.

3 Wanli, X. et al. (2015) 'Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory', Computers in Human Behavior, Elsevier, 47, pp. 168-181. doi: 10.1016/j.chb.2014.09.034.