Wu, Y.-T., Chang, M., Li, B., Chan, T.-W., Kong, S. C., Lin, H.-C.-K., Chu, H.-C., Jan, M., Lee, M.-H., Dong, Y., Tse, K. H., Wong, T. L., & Li, P. (Eds.). (2016). *Conference Proceedings of the 20th Global Chinese Conference on Computers in Education 2016*. Hong Kong: The Hong Kong Institute of Education.

Educational Data Mining using Chance Discovery from Discussion Board

Simon, Y. K. LI [#], Gary, K. W. WONG [*]

The Hong Kong Institute of Education, Hong Kong

[#] liyk@ied.edu.hk

* wongkawai@ied.edu.hk

Abstract: *We present our preliminary research results of identifying serendipitous findings from discussion forums of students in this paper. A text-mining tool called Polaris was used to generate the analytical results in KeyGraph helping teachers assess the effectiveness of teaching and innovation of learning respectively through the visualization of hidden patterns in the online learning environment. Our results show that the serendipitous findings have shown a traceable pattern to predict the academic performance of students. The research findings can lead to adaptive pedagogical designs for teaching and learning by finding hidden patterns and linkages among the students' serendipitous learning. Both teachers and students can be supported on how to improve teaching and learning with feedbacks from this new tool. Ultimately, the advanced mining technology to assess the students' knowledge discovery process can create a new approach for transformative learning and teaching in education.*

Keywords: educational data mining, topic detection, serendipitous learning, chance discovery, learning analytics, big data.

1. Introduction

E-learning systems have become more popular recently in the higher education sector producing a large volume of data created by students and teachers inside via collaborative online discussion forums. Therefore, teachers are difficult to analyze the data without the help of data mining technologies know as *learning analytics.* It helps to collect large-scale of data, extract actionable patterns, and obtain insightful knowledge from educational data so that educators can better understand the thinking patterns of students during the learning process, as well as capture and visualize their behavioral intention or motivation in learning. This information will facilitate the assessment of students through a systematic and real-time approach, to identify effective pedagogic changes for particular students, and to guide them through the learning process with the ultimate goal of optimizing their learning outcome. There has been a limited research work but a growing interest in how *serendipity*, as a by-product of this collaborative online learning, can occur among the students through the participation in the online environment. This research can open a new window to this new exciting area to extract innovative ideas from discussion of both students and teachers.

2. Project Objectives

This project aims to develop innovative analytical methods and tools facilitating formative assessment of teachers to evaluate the students learning in the online learning environment through the concept of serendipity to determine if students are able to generate new knowledge beyond the expected learning activities. Teachers can therefore assess more effectively if students are able to achieve the Course Intended Learning Outcomes (CILOs) and some of the Generic Intended Learning Outcomes (GILOs) in the Institute, such as creative thinking skills and global perspectives. The outcomes of this project can inform the Institute expectation in developing diversified and innovative assessment methods with advanced e-learning technology.

3. Prior Works Used by This Project

3.1. KeyGraph

The KeyGraph algorithm analyzes a document by using an analogy of physical building construction metaphor for the document. The KeyGraph algorithm first produces co-occurrence graphs of keyword items from the document into different clusters. Afterwards, the relationships between items in the clusters and amongst the clusters would be extracted. so that the most important ideas would appear. The top ranked terms based on each term's relationship to the clusters are selected as keywords. Each cluster represents a concept / idea from the author of the document. The most frequently found items in the document would be depicted as black nodes in KeyGraph. The most strongly co-occurring item pairs are connected by black solid lines as the edges in the KeyGraph representing the co-occurrence between events. As a result, these extracted terms can accurately match the main ideas of the document in finding opportunities.

3.2. Chance discovery (CD)

Chance discovery (CD), a concept and method invented by Professor Yuiko Ohsawa, would also be adopted in this project to detect some "less explored" or even "unexplored" areas by using a text-mining algorithm called KeyGraph (Ohsawa et al., 1998). CD is the suite of tools helping to find the possible "signs" of chance, which may be unexpected relevance (Ohsawa, 2005). A "chance" is described as a rare, hidden, unnoticed, potential, novel, but significant event(s)/situation(s) that can be conceived either as a future opportunity or risk (Ohsawa & Fukuda, 2002). A chance can therefore be defined as a piece of timely information about an event or a situation having significant impact on decision-making. By interactively viewing the results using review workshops in the Innovator Market Game (IMG) (Wang, Ohsawa, & Nishihara, 2011), reviewers would meet together to evaluate the possible "chances" from KeyGraphs generated. Once those innovative ideas are found, pairwise comparison using Analytic Hierarchy Process (AHP) would then be performed for each criterion with each innovative idea by using different weight factors forming a matrix calculation with hierarchical weights being assigned to ensure each criterion would thoroughly be considered for each innovative idea to produce a fair judgment.

4. Experiment Design and Preliminary Results

In this project, a total 24 undergraduate students in the Hong Kong Institute of Education from the General Education course called "Technology, Entertainment and Mathematics" have been sampled for this preliminary experiment. These students taken this free elective of general education course were ranging from Year 2 to Year 4 in their 4-year of study. One of the course requirements was to complete a reflective posting on an online discussion forum in Schoology. They were asked to watch a BBC documentary film called "Beautiful Equations" and other selected movies. Afterwards they posted their reflections in the forum. Each student was also required to comment on three self-selected peers, which were extracted in our experiment as text files for analysis. Polaris (http://www.panda.sys.t.u-tokyo.ac.jp/KeyGraph/), a software tool called developed by Oshawa Lab, was used to visualize KeyGraphs by inputting the extracted text files. The software tool can construct KeyGraphs in terms of black (i.e. keywords making some extracted concepts) and red (i.e. keywords that potentially connect multiple concepts) nodes, which might be regarded as "chances". There were a total of 110 ("Beautiful Equations") and 94 (Selected Movie) posts in the forums from 24 students who had completed the related study module were sampled. The preliminary results from using KeyGraph analysis were displayed in Fig. 1 and Fig. 2 below. The collective traits of understanding the main themes of the film could be found from the keywords shown below. The results can indicate that the general appreciation of beautiful mathematics has resulted. Fig. 2 also indicates some similar patterns of appreciation of mathematics from the keywords extracted.

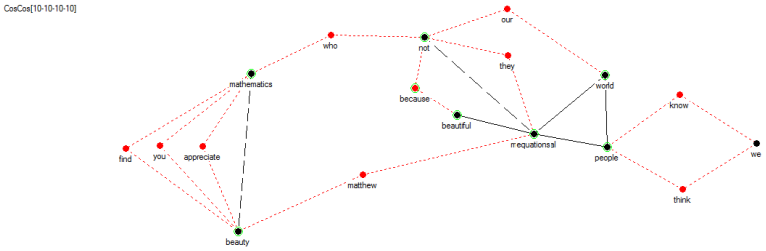

Figure 1. Output from KeyGraph showing key concepts and chances ("Beautiful Equations")

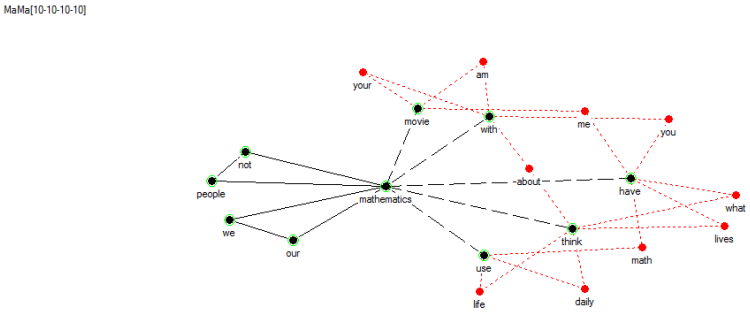

Figure 2. Output from KeyGraph showing key concepts and chances ("Selected Movie")

5. Potential Contributions to Learning and Teaching

5.1. Formative assessment and learning analytics

With advanced data analysis, feedback from the assessment can be given instantly with prediction as well (Tempelaar, Rienties, & Giesbers, 2015). By using these data mining techniques to effectively analyse the interaction data and systems information and academic information collected from e-learning systems, teachers can better understand the deep thinking and motivation of students during the learning process (Mazza & Milani, 2004; Romero et al., 2008). The assessment of students can be facilitated through a systematic and real-time approach with effective pedagogic changes for particular students (West, 2012) through the learning process to optimize their learning outcome (Ferguson, 2012).

5.2. Research studies in serendipitous learning

Recently, research begins to study how "serendipity" plays an important role in information seeking (Foster & Ford, 2003; Eynon, 2013), and there seems a huge potential of unexpected learning discoveries in collaborative online learning (Buchem, 2011; Zhang, Liu, & Si, 2011). Existing works indicate a potential connection of serendipitous learning to students' online learning process as "occurring when hidden connections or analogies are unexpectedly discovered, mostly during searching processes which are typical for informal learning activities" (Gaeta et al., 2014). This concept of "serendipity" in the learning analytics not yet been fully addressed by the existing algorithms, and research in this area remains an open question (Kop, 2012; Shum, & Ferguson, 2012; Slade, & Galpin, 2012; Delgado Kloos et al., 2014).

6. Conclusions

In this project, we present the opportunities to use the data mining technologies for facilitating learning analytics by performing text mining over the unstructured online discussion data. The preliminary results are stimulating for further investigation into deeper understanding of the results with different perspectives discussed in the next section.

7. Future works and Discussion

More different perspectives and variables together with their constraints would be explored in the coming stages of our research project such as the sample sizes, multiple academic subjects, writing styles, and other qualitative factors. We would proceed to the stage two and three of CD to increase the accuracy of the experiments with human inspection and evaluation. Furthermore, semantic technologies would be considered to understand the keywords and their association.

Acknowledgements

This project is funded by the Teaching Development Grant of The Hong Kong Institute of Education (Ref: TDG 2015/16).

References

Buchem, I. (2011). Serendipitous learning: Recognizing and fostering the potential of microblogging. *Form@ re-Open Journal per la formazione in rete*, *11*(74), 7-16.

Delgado Kloos, C., Ibanez-Espiga, M. B., Fernández-Panadero, C., Munoz-Merino, P. J., Estevez-Ayres, I., Crespo-Garcia, R. M., & Perez-Sanagustin, M. (2014, October). A multidimensional analysis of trends in educational technology. In *Frontiers in Education Conference (FIE), 2014 IEEE* (pp. 1-4). IEEE.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5), 304-317.

Gaeta, M., Loia, V., Mangione, G. R., Miranda, S., & Orciuoli, F. (2014). Unlocking serendipitous learning by means of social Semantic web. In *CSEDU 2014-In proceedings of the 6th International Conference on Computer Supported Education* (Vol. 1, pp. 285-292).

Kop, R. (2012). The unexpected connection: Serendipity and human mediation in networked learning. *Journal of Educational Technology & Society*, *15*(2), 2-11.

Mazza, R., & Milani, C. (2004, November). Gismo: a graphical interactive student monitoring tool for course management systems. In *TEL'04 Technology Enhanced Learning'04 International Conference* (pp. 18-19).

Ohsawa, Y. (2005). Data crystallization: chance discovery extended for dealing with unobservable events. *New Mathematics And Natural Computation*, *1*(03), 373-392.

Ohsawa, Y., Benson, N. E., & Yachida, M. (1998, April). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on* (pp. 12-18). IEEE.

Ohsawa, Y., & Fukuda, H. (2002). Chance discovery by stimulated groups of people. Application to understanding consumption of rare food. *Journal of Contingencies and Crisis Management*, *10*(3), 129-138.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

Slade, S., & Galpin, F. (2012, April). Learning analytics and higher education: ethical perspectives. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 16-17). ACM.

Shum, S. B., & Ferguson, R. (2012). Social learning analytics. *Journal of Educational Technology & Society*, *15*(3), 3-26.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning Analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157-167.

Wang, H., Ohsawa, Y., & Nishihara, Y. (2011). A system method to elicit innovative knowledge based on chance discovery for innovative product design. *Multidisciplinary Studies in Knowledge and Systems Science*, 153.

West, D. M. (2012). Big data for education: Data mining, data analytics, and web dashboards. *Governance Studies at Brookings*, 1-10.