Data Mining in Educational Assessment: A Perspective of Big Data

Apitchaka Singjai
Chiang Mai University, College of Art Media and Technology, Software Engineering Department, Chiangmai, Thailand,
apitchaka@camt.info

Abstract

At the present time, data mining is a domain of enormous potential representing a great challenge for researchers. On the other hand, competency management is a discipline with growing importance, providing an important tool of decision-making for the management. Our paper investigates the potential of big data analysis from a competency management perspective regarding educational organizations. Competency in educational organizations need to be properly assessed and developed to provide efficiency. The management of educational organizations should be supported properly in the decision-making process in order to guarantee the right performance. Data mining techniques could extract the hidden knowledge of organizational data, providing valuable inputs for decision-making and potentially enhancing the quality of educational systems. The tendency is that there is a fast growing amount of data generated in educational organizations making data mining techniques a crucial element of managerial support. What are the key issues of managing competencies in educational organizations and how can data mining techniques be implemented in this field? This is the fundamental research question behind our paper.

Introduction

Big data refers to the great amount of data that is available for the organization including the new data that can be discovered during the data analyzing (Huebner).

In the big data era, the large dataset tends to be an important part of business because of the long tradition in natural sciences (Steiner et al., 2014). There are two main factors to be mentioned; firstly, understanding how technology influences and impacts the assessment type and process; secondly, due to a developing confidence in creating and analyzing argument from evidence, big data has become more significant than ever (Gibson, 2015). Not only in business sectors but also in educational sectors.

The educational institutions should provide up-to-date information about the institutional effectiveness (C. Romero and Ventura, 2010). According to the reason stated before, the data analysis is required. Moreover, the learning activities let the educational institutions deal with large amounts of data (Ferguson,2012).

Data mining could enhance the decision making by analyzing the hidden patterns and relationships among big data (Huebner, 2013). There are several sectors where data mining has been successfully applied such as industries, governmental sector, military, retail and banking, but typically not in the educational sector. (Ranjan and Malik, 2007).

The institution needs data mining to reveal the significant data for decision making (Kiron et. al., 2012). Some data should be ignored to let the institution's goals come closer (Slade and Preinsloo, 2013). Large amount of learners come with large data, the institutions themselves require budgets and need to focus on quality and accountability (Macfadyen & Dawson, 2012). Day by day, data is generated from various sources, so that the educational

ISBN 978-602-19725-8-8

institution requires the technological solution to deal with the data challenge. (Macfadyen et al., 2014).
According to the institution's requirements, academic analytics should be applied. Academic analytics also represent the sub-field of educational data mining. The term educational data mining is wider than the academic analytics because the educational data mining focuses on any type of data in institutions. Academic analytics applies a macro perspective that focuses on the processes from department level up to the university level related to institutional effectiveness and student retention issues, ignoring details of individual courses. (Huebner, 2013).

Data Mining Techniques for Education

In educational data mining, data mining tools and techniques are applied to improve the learning experience and institutional effectiveness (Baker and Yacef,2009 ; Huebner, 2013). The key techniques that are used for education related data are web mining, classification, associated rule mining and multivariate statistics (Calders and Pechenizkiy, 2012)

Web mining

Web mining is one of the data mining techniques that deals with web data including web content, web structure and web usage data (Srivastava et al., 2005). The web mining research committee also introduced key concepts of web mining as follows:

1.1.1. Ranking Metrics: the metric can rank the quality of pages if they are relevant.
1.1.2. Robot Detection and Filtering: web robots are software programs that can separate the robot behavior from the human behavior.
1.1.3. Information Scent: information scent concept deals with the application of the relevant information for browsing.
1.1.4. User Profile: user profile is related to the decision making process for limiting access to websites.
1.1.5. Interestingness measures: the interactions between authors and readers are mapped and tracked..
1.1.6. Preprocessing: preprocessing is about preparing the data for mining.
1.1.7. Identifying Web Communities of Information Sources: the communities are a hub to link the user and the information source together.
1.1.8. Online Bibiliometrics: online repositories of publications make the access of papers easier for researchers.
1.1.9. Visualization of the World Wide Web: the visualization tool can support the web data analysis.

If the institutions propose online assessment, the web mining technique should be adopted. Izso and Toth applied web-mining methods to analyze the student behavior in a virtual learning environment course. The server provided log files about the interactions between learners and the electronic syllabus during the course. Parts of the empirical results have been published (2008).

ISBN 978-602-19725-8-8

Classification

Classification is one of the useful traditional data mining techniques which is commonly applied in e-learning by developing the model that can classify the great amount of data from the pre-classified dataset. (Ahmed and Elaraby, 2014; Bhardwaj and Pal, 2011). The classification is a predictive data mining technique that predicts the unknown value of data by using different known value of data (Radaideh et. al., 2006). Predictive modeling is used to map vector input to the scalar output in measurements (David et. al., 2001). In another words, classification is about mapping the data into predefined groups of classes (Bhardwaj and Pal, 2011).There are two processes involved in data classification (Bhardwaj and Pal, 2011).

1.1.10. Learning: the data is analyzed by classification

1.1.11. Classification: the accuracy of test data is estimated under the classification rule.

Before applying the data mining technique, the predefined groups of classes should be known to initiate the value of data. Ahmed and Elaraby presented in their work an analysis of data from the attributes to predict the final grade of the students. The classification method that they applied is a decision tree with ID3 algorithm (2014).

Association Rule mining

Association analysis can discover the frequency conditions from the given database to make association rules (Ahmed and Elaraby, 2014).
In data mining, association rule mining is used to identify concrete rules to discover the relationship between the variables in large data by using the different measurements of interest (Bambrah et. al., 2014). There are two problems associated with rule mining. The first problem is to search for a large dataset that exceeds the predefined threshold. The second problem is setting up the association rules from the large dataset. Associated rule generating is the main task of the association rule mining technique.
In a study from 2014 researchers adopted associated rule mining technique to enhance the performance of students. They utilized an a priori algorithm to extract the educational database and then specified and classified it based on the student performance. The types of data represented were as follows: data on assignments, internal assessment, and attendance. Three types of students were represented as regards the level of performance: poor, good, and excellent students. By processing the the report results, the students could improve their performance in the following semester (Bambrah et al., 2014).

Multivariate statistics

Multivariate statistic is about clustering the common data together. In multivariate approach, feature selection is applied to reduce the dimensionality (Sillin et. al., 2014). When there are a lot of factors to make a decision, the multivariate problem is occurred. Popular techniques for solving multivariate problems are principal component analysis, cluster analysis and discriminant analysis 17 (Wu et al., 2014). Clustering allows finding the

ISBN 978-602-19725-8-8

group for the data. Cluster analysis is the technique that groups samples with similar characteristics, while the principal component analysis is a technique that transforms the data in the dataset into a simple one (Cloutier et al. 2008). If there is a variety of decision-making factors, multivariate statistics technique is an appropriate tool to minimize the problem of sophisticate analysis.
A group of researchers have recently aimed to classify education systems into 2 categories which are: exceptional and fair. They analyzed the educational systems from 64 countries with three different types of multivariate statistics techniques as follows: principle component analysis, factor analysis, and discriminant analysis. As a result, the researchers found that the principle component analysis is the most appropriate tool among the three mentioned above because it could reduce the dimensionality from 20 variables into five (Brooks et. al, 2015).

Educational Assessment

The challenge of big data in educational assessments composes of 4 main factors: time sensitivity, the digital performance problem space relationship, the layer of aggregation and translation, and the representation of dynamics (Gibson, 2015).

Time Sensitivity

Time sensitivity is about how to deal with change over time with respect to data flow. Data should be understood as patterns. Simultaneous and sequential interactions are required. More and more data is being produced because time-based data is increasing every minute. In the big data era, the amount of data is not so interesting from a research perspective but the changing in data over time.

Digital Performance Relationship

Nowadays, the educational activities mostly rely on the internet in terms of online systems. The interactions among digital performance space relationships represent an important domain of educational assessment. This interaction includes actions, communication, and products. Users can create own digital space to empower the technology adaptation. The digital performance space relationship can enhance the ability of complexity in performance settings.

Layers of Aggregations and Translations

The abstract data layers are to define which part of the whole dataset the single data belongs to. The main objective of these layers of interpretation is to form groups of data into meaningful datasets. Translation is a correlation relationship related to the trend of configuration information by the duplication and consequences, while aggregation is a relationship from the event of the algorithm that rely on criteria to compute (Debar and Wespi, 2001).

Presentation of Dynamics

ISBN 978-602-19725-8-8

The presentation of dynamics is about the representation of interaction in digital performance spaces. The performance should be scored as a summative assessment (Newhouse, 2014). There are 2 layers for dynamics representations such as qualitative and quantitative representations (Tafani et. al., 2002). The translation process should be in electronic form rather than in manual. Every data can be reconstructed and transformed into a visualized form.

The Table of Recommendations

Data-driven decision making becomes more sophisticated when implementing the concept of big data which refers to software-based analytics (Picciano, 2012). Especially when taking assessment aspects into consideration, proper techniques are required to analyze great amount of data because not only do we have to deal with the available data, but also with new data generated by the process of analysis. The table below shows the relationship between the data mining techniques and educational assessment issues.

Table 1: Table of recommendation: data mining techniques from an educational assessment point of view

	Web Mining	Classification	Associated Rule Mining	Multivariate Statistics
Time Sensitivity	-	-	/	-
Digital Performance Relationship	/	-	-	-
Layers of Aggregation and Translation	-	/	-	/
Presentation of Dynamics	/	-	/	-

Table1 illustrates the possibilities of data mining technique applications in the domain of educational assessment. As described before, various data mining techniques have various features and various potentials that define their specific domains of application. On the other hand, the educational assessment is an area with 4 clearly differentiated factors. Given this diverse nature of the assessment challenges, and the variety of data mining techniques, our research paper proposes to demonstrate a recommendation table where techniques are assigned to the various assessment issues. There are 4 data mining techniques in the table which are web mining, classification, associated rule mining, and multivariate statistics. The 4 factors of educational assessment represented are time sensitivity, digital performance relationship, layers of aggregation and translation, and presentation of dynamics. Firstly, associate rule mining is recommended to be applied for time sensitivity related problems. The big amount of data form date and time should be prioritized and analyzed using the associate rules. Secondly, on the domain of digital performance relationship the web mining technique should be adopted to elicit the performance from the digital form. Thirdly, the grouping through classification techniques and multivariate techniques can be adopted for layers of aggregation and translation. Lastly, the presentation of dynamics, which is about presenting interactions of the digital performance space, can be supported by web mining in terms of using online systems to obtain the data and can also be supported by associate rule mining to analyze the data.

ISBN 978-602-19725-8-8

Conclusion

Assessment is an important tool for them to enhance the performance of the educational organization. The assessment should incorporate and process the data to support the decision-making process. The great amounts of data in educational organizations require various data mining techniques. Also, the various dimensions of educational assessment reveal characteristic challenges. This paper gave an overview of data mining techniques and assessment challenges and proposed a recommendation table that assigns specified techniques to assessment tasks taking both the nature of the specific task and the characteristics of the given technique into consideration. The four techniques characterized in this paper are: data mining technique, classification, associated rule mining, and multivariate statistics. These techniques are assigned to the following educational assessment domains: time sensitivity, digital performance relationship, layers of aggregation and translation, and presentation of dynamics.

Reference

Huebner, R. A. (2013). "A Survey of Educational Data-mining Research." *Research in Higher Education Journal.*

Romero, C. and S. Ventura (2010). "Educational Data Mining: A Review of the State of the Art." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 40(6): 601-618.

Ranjan J. and K. Malik (2007). "Effective educational process: a data mining approach." *VINE* 37(4): 502-515.

Baker, R. and K. Yacef (2009). "The State of Educational Data mining in 2009: A Review and Future Vision." *Journal of Educational Data Mining* 1(1).

David Kiron, Rebecca Shockley, et al. (2012). "Analytics: The Widening Divide." *MIT Sloan Management Review* 53(2): 1-22.v

Calders, T. and M. Pechenizkiy (2012). "Introduction to the special section on educational data mining." SIGKDD Explor. Newsl. 13(2): 3-6.

Gibson, D. (2015),"Big Data in Educational Assessments", available at http://www.unescobkk.org/education/ict/online-resources/databases/ict-in-education-database/item/article/big-data-in-educational-assessments/ (assessed 12 Febuary 2015)

Macfadyen, L. P. and S. Dawson (2012). "Numbers are not enough: Why e–learning analytics failed to inform an institutional strategic plan." *Educational Technology & Society* 15(3): 149-163.

Slade, S. and P. Prinsloo (2013). "Learning analytics: Ethical issues and dilemmas." *American Behavioral Scientist* 57(10): 1510-1529.

Leah P. Macfadyen, Shane Dawson, et al. (2014). "Embracing Big Data in Complex Educational Systems: The Learning Analytics Imperative and the Policy Challenge." *Research and Practice in Assessment* 9.

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

ISBN 978-602-19725-8-8

Christina M. Steiner, Michael D. Kickmeier-Rust, et al. (2014). Learning Analytics and Educational Data Mining: An Overview of Recent Techniques. *Learning Analytics for and in Serious Games*.

T. Srivastava, P. Desikan, and V. Kumar, *Web Mining – Concepts, Applications and Research Directions*. Studies in Fuzziness and Soft Computing, ed. F.a.A.i.D. Mining. Vol. 180. 2005: Springer Berlin Heidelberg. T. Srivastava, P. Desikan, et al. (2005). *Web Mining – Concepts, Applications and Research Directions*, Springer Berlin Heidelberg.

Lajos Izsó and P. Tóth (2008). "Applying Web-Mining Methods for Analysis of Student Behaviour in VLE Courses " *Acta Polytechnica Hungarica* 5(4): 79-92.

Brijesh Kumar Bhardwaj and S. Pal (2011). "Data Mining: A prediction for performance improvement using classification." *International Journal of Computer Science and Information Security* 9(4).

Abeer Badr El Din Ahmed and I. S. Elaraby (2014). "Data Mining: A prediction for Student's Performance Using Classification Method." *World Journal of Computer Application and Technology* 2(2): 43-47.

Bambrah, C., M. B. , et al. (2014). "Mining Association Rules in Student Assessment Data." *International Journal of Advanced Research in Computer and Communication Engineering* 3(3): 5340-5342.

Jianhua Wu, Peiyue Li, et al. (2014). "Using correlation and multivariate statistical analysis to identify hydrogeochemical processes affecting the major ion chemistry of waters: a case study in Laoheba phosphorite mine in Sichuan, China." *Arabian Journal of Geosciences* 7(10): 3973-3982.

Cloutier V, Lefebvre R, et al. (2008). "Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system." *Journal of Hydrology* 353(3-4): 294-313.

Al-Radaideh, Q. A., E. M. Al-Shawakfa, et al. (2006). Mining Student Data using Decision Trees. *International Arab Conference on Information Technology*. Jordan.

David Hand, Heikki Mannila, et al. (2001). *Principles of Data Mining*, A Bradford Book The MIT Press.

Brett Vaughan, et al., *Methods of assessment used by osteopathic educational institutions.* International Journal of Osteopathic Medicine 2012. 15: p. 134-151. Brett Vaughan, Vivienne Sullivan , et al. (2012). "Methods of assessment used by osteopathic educational institutions." *International Journal of Osteopathic Medicine* 15: 134-151.

Picciano, A. G. (2012). "The Evolution of Big Data and Learning Analytics in American Higher Education." *Journal of Asynchronous Learning Networks* 16(3): 9-20.

NEWHOUSE, C. P. (2014). "Using Digital Portfolios for High-Stakes Assessment in Visual Arts." *Research and Practice in Technology Enhanced Learning* 9(3): 475-492.

Ashley Brooks, Amber Shoecraft, et al. (n.d.). "Education By Nation: A Multivariate Statistical Analysis " , available at http://www.units.miamioh.edu/sumsri/sumj/2006/Stat-Education%20Paper.pdf. (accessed12 Febuary 2014)

Hervé Debar and A. Wespi (2001). *Aggregation and Correlation of Intrusion-Detection* Alerts, Springer Berlin Heidelberg.

Tafani, E., S. Bellon, et al. (2002). "The role of self-esteem in the dynamics of social representations of higher education: An experimental approach." Swiss Journal of Psychology 61(3).

ISBN 978-602-19725-8-8