BIG DATA ANALYSIS IN EDUCATION

Manaswini Arcot

Department of Computer Science, Villanova University

INTRODUCTION

As most of our commerce, entertainment, communication, and learning are occurring over the Web, the amount of data generated by online activities is increasing rapidly. According to a report by Association for the Advancement of Computing in Education on Big Data and Education: Prospects and Problems, a Statement by Stefanie Panke, “many education reformers see the merging of student data, predictive analytics, processing tools, and technology-based instruction as the key to the future of education and a way to further opportunities and equity in education.”

This Paper briefly describes the learning analytics and the educational datamining techniques applied on educational data. Data mining supports a variety of education-related functions, such as building student models to individualize instruction, evaluate pedagogical support, and contribute to learning science. [1]. Section I gives an overview of the applications of big data in learning. Section II gives detailed discussion of big data in eLearning, Sub sections of section II deals with Big data in terms of e-learning. B) Benefits by analyzing the data using big data techniques. Section III deals with Big data in Higher Education, sub sections are III A) Brief Introduction in higher education B) technical challenges by the researchers to analyze the data C) Example 1 D) Example 2 gives the overview about the challenges faced and the prediction models used in the research and the final Section IV includes the Conclusion.

SECTION I: APPLICATIONS OF BIG\_DATA IN LEARNING:

Big Data techniques can be used in enormous ways in learning analytics as listed below:

* Performance Prediction of Student's
* Attrition Risk Detection
* Data Visualization
* Intelligent feedback
* Course Recommendation: New courses can be suggested to students based on their interests by analyzing their activities. That will guide students in choosing fields in which they may not have interest.
* Student skill estimation
* Behavior Detection: Detection of student behaviors in activities organized by communities or games which helps in developing a student model
* Grouping & collaboration of students
* Social network analysis
* Developing concept maps
* Constructing courseware
* Planning and scheduling [13]

Section II: BIG DATA IN E-LEARNING:

E-learning is education through the Internet, network, or standalone computer. ELearning is constantly evolving in today’s World. The emerging technologies and improved instructional design models, the potential outcomes are huge for the future of eLearning. [2]

II.A) Big data, in terms of the eLearning:

It is the information that is made by learners while they are taking an eLearning course or training module. This type of data is collected by the Learning Management System (LMS), the eLearning Authoring Tool, social media, multimedia, etc., that are set by the organization or the eLearning professionals. LMS is a software application that provide complete course-delivery functions—administration, documentation, content assembly and delivery, tracking and reporting of progress, user management and self-services, etc. LMS are Web based and are considered as a platform to build and deliver modules and courses. Open-source examples include Moodle, Sakai, and ILIAS. [9]. There can be individual pieces of data that can be collected and analyzed to decide how the learner is securing data, at what pace, and to pinpoint any issues that may exist inside the eLearning procedure itself. [3]. For example, if an employee of an organization is working on a training module based on the company policies, his/her progress, assessment results, social sharing, and any other data being generated during the eLearning course is “big data”.

II. B) Benefits by analyzing the data using Big data techniques:

1. Permits eLearning experts to see how the learners are processing the data and which adapting needs offer the most to them.
2. Empowers eLearning experts to pinpoint regions that may should be calibrated inside the eLearning course or module.
3. Gives an analysis of which eLearning modules are shared most.
4. Data is received almost immediately, rather than waiting for long time to collect the assessments. This implies that eLearning experts can start executing changes or using the information to adjust their eLearning methodology immediately.
5. Based upon patterns, eLearning experts have the ability to foresee where learners may struggle or exceed expectations.

For instance, if a learner is provided with an analysis of where he/she faced the problem, while taking the eLearning course, they can then figure out how to correct the issue in future. At the same time, if the online trainer observes that the most of the learners struggle with a specific module or assignment, they can amend the course structure to improve learners’ performance. Big data analysis gives us an insight of which strategies are working and which are not, and also allows to determine which eLearning courses are helpful to develop skills and which eLearning modules or elements may be inapplicable. [3].

According to an eLearning article, the learning management system like *Lectora Express* – *The Easy LMS* or *Course Mill LMS* helps to track, record and analyze the learners’ Big Data. [4]. Pictorial representation of the Students Progress through bar graphs and histograms representing the student’s problems and particular activities and a tree structure to represent the knowledge, can increase the interaction levels between the student and teacher and helps the student to improvise his skills.

Section III: BIG DATA IN HIGHER EDUCATION:

III.A) Brief Introduction:

Many Universities and Colleges have indicated that big data analytics for an institution, provides them an idea to improve their standards and services and also help them in competing with the other institutions. [7]. Now a Days in Higher education, it is important to anticipate student’s performance. There are two reasons for this: it is a key to distinguish which set of students would do well in the semester end examination with the goal that they can be awarded grants and more vitally to recognize the students who may fall short at the end in examinations so that some type of remediation may be offered to them. [6]. Currently, educational data mining is looking forward to focus on developing new tools to discover patterns in data. The U.S. Department of Education’s National Education Technology Plan, as one part of its model for 21st-century learning powered by technology, visualizes the way of using data from online learning systems to improve instruction. [9].

Factors that affect the strong applications in this field are the cost and the challenges. The cost associated in collecting the data and storing the logged data is handled by the IT departments and the algorithm developers calculates the computational costs, that is required to implement these techniques.

III.B) Technical Challenges:

Involved are the data from the educational data systems cannot be exchanged. Correlating the administrative data and the Classroom level data is a big challenge that is faced by the researchers. Combining this data will be more helpful for the algorithms to predict the results on the student data. [9]

B. a) Factors/Features Considered for the Analysis:

The academic performance of students depends on several factors; some of these are previous academic records, economic status, family background, performance in mid semester examinations, participation in forums etc. From these data, the benefits of analytics would be in the areas of student’s recruitment and retention and student’s performance. [7]. The U.S. Department of Education collected data from Web pages and unpublished documents on learning analytics and educational data mining.

Important features are time, sequence, and context. [9]. Time is used to calculate the length for which the session is being opened, and the time taken by the students to learn. Sequence represents how concepts build on one another and how the practice sessions and the tutorial sessions should be organized. Context is important for explaining results and knowing where a model may or may not work. [9]. Researchers found that, considering the assignments, concept strands, standards, and students who do or do not have mastery of the concepts in a standard is more productive than classifying students into groups based on learner types. In contrast, others also classify students based on understandings, learning trajectories, motivation, and cultural background.

III.C) EXAMPLE -1

According to the Research that was conducted by Bharathidasan University, Tiruchirappalli, India on their own college student’s data, predicted the Results of the students. Their data contains student details of different subject marks in semester wise, that have been recorded and subjected to the data mining process. Data Processing methods used are the Data Clustering. Clustering is the process of dividing data into groups of similar objects. K-means algorithms is one of the best algorithms of Clustering and this was used by the researchers of this university. To build a model, a complete system is required that states a problem, identifying the data that is required and solutions to the problem. For this they have conducted interviews with domain experts to get the information required to solve the problem. With the data collected, they have extracted the required knowledge using data mining techniques. And they have applied the Naïve Bayes Probabilistic Algorithm. They used the WEKA Software, which is a Java based open source tool developed by the University of Waikato in New Zealand to predict the pass percentage and fail percentage of the Overall students appeared for a particular examination. They have chosen sample of 38 students record for the analysis, they have predicted the results that 36 will pass and one absentee and the one failure out of the records. [10]. This shows that by applying the proper data mining techniques on student’s performance data, we can obtain a lot of hidden knowledge / information from the vast data, which can be further used by the management of the educational institutions in decision making process.

III. D) EXAMPLE-2

Another Example from a case study by Srivatsan Ramanujam, a consulting project was started by Pivotal Data Labs which included the Apache Hadoop and Pivotal Greenplum Database. [11]. The data sources that have been collected for the analysis was from four systems: a) Online Applications for Education: This included all the web log data from the online applications and the class assignments and work submissions b) Forums: The Education institutions also had the dashboards, the discussion boards which contains the student questions, answers and their views. c) Help Desk: The place where students raise tickets for the questions and this also contains the time stamps and the topics and categories d) Student Demographic and Operational Information: This contains demographic, courses, age, background, test scores, GPA, previous education, applications, admissions, and enrollments.

D. a) Challenges faced:

The source data contains the personally identifiable information. Agreements with laws and regulations, like HIPAA and FERPA. They have masked the PII of each data source that was non-reversible masking and few PII were removed. Finally, the masked data was entered into the database that allowed the data scientists to work on the anonymized data. [11].

D. b) PREDICTION MODELS:

They have developed models on the segmenting the student data. they have used the Madlib’s

K –means Algorithms to segment the data into clusters based on different features like the demographics and the discussion board participation. There were two types of predictions that Pivotal labs used --- one was for predicting student retention and the other for predicting success in their grades with timely graduation. Madlib’s linear regression and elastic net regression models were used to build success prediction models, revealing what good student profiles looks like based on the data. They found few features that helps the students at risk to enable them to succeed.

Other Prediction model was using the Madlib’s logistic regression model, to predict student retention, discovered the data characteristics for students who leave or complete their degree. This helped as an early warning system that detects the students profile which matches this behavioral profile. After understanding the behavior and predictors for the retention and success, the team realized that these groups can have impact on the other groups. For example, students at low risk can help the students who are at risk. The project helps the educational in determining the cause and effect. [11]

III.E).EXAMPLE - 3

A data mining research was performed at a Bulgarian university, the research work was proposed to find the patterns if any, in the data available, that may be used in predicting students’ performance at the university, based on their personal and previous university data. [12]. Their project goal was to classify university students, based on their performance and the pre-university characteristics. They collected data related to the personal data of university applicants (names, addresses, secondary education scores, selected admission exams, etc.), data about the organization and performance of the admission exams, scores achieved by the applicants at the admission exams. All the data related to the student performance at the university which includes student personal and administrative data, the grades achieved at the exams on the different subjects., etc. The data described 20 different parameters. Along with this data they additionally have the data of the student’s previous university characteristics, that include place and profile of the secondary school, the final score, the score achieved at admission exam, and the total admission score. Based on the university score achieved by the students they classified the students in 5 distinct categories. They applied the model on 10330 instances. They have used the open source tool WEKA for the project implementation.

SECTOIN IV: CONCLUSION

Big Data, will continue to affect the future of e-learning by creating wide range of opportunities for online training to be more effective. Better understanding of challenges from user’s point of view by education researchers and IT developers, for implanting and adopting analytics and data mining techniques in classrooms schools, districts, and other institutions to be successful

REFERENCES:

[1] <http://blog.aace.org/2014/12/22/big-data-education-prospects-problems/>

[2] <http://elearningindustry.com/big-data-in-elearning-future-of-elearning-industry>

[3] <http://elearningindustry.com/big-data-in-elearning-future-of-elearning-industry>

[4] <http://elearningindustry.com/big-data-and-how-it-is-changing-e-learning>

[5] <http://ijarcst.com/doc/vol3issue1/ver1/ahmed.pdf>

[6] <http://research.ijcaonline.org/volume107/number1/pxc3899939.pdf>

[7] <http://net.educause.edu/ir/library/pdf/ers1207/ers1207.pdf>

[8] <http://www.wgu.edu/admissions/academic_experience_aap>

[9] <https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf>

[10] <http://www.ijcsit.com/docs/Volume%205/vol5issue04/ijcsit20140504249.pdf>

[11]<http://blog.pivotal.io/data-science-pivotal/case-studies/big-data-in-education-analyzing-student-clusters-to-influence-success-and-retention>

[12] <http://www.csie.ncue.edu.tw/~myshih/103-1/54019/Kabakchieva.pdf>

[13] <http://ictactjournals.in/paper/IJSC_Paper_6_pp_1035_1049.pdf>