International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-4)

Research Article | **April 2017**

Data Mining Model for Effective Data Analysis of Higher Education Students Using MapReduce

Shankar M. Patil*
Research Scholar, Department of Computer Engineering,
Singhnia University, Rajsthan, India

Dr. Praveen Kumar
Professor, Department of Computer Engineering,
Meerut Institue of Engg. & Technology, U.P., India

Abstract—

Data analysis plays an important role for decision support irrespective of type of industry like any manufacturing unit and educations system. There are many domains in which data mining techniques plays an important role. Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. In this paper we used educational data mining to improve graduate students' performance, and overcome the problem of strong and weak of graduate students. In our case study we try to extract useful knowledge from graduate students data collected from the BV's College of Engineering . The data include four years of period [2012-2016]. After preprocessing the data, we applied data mining techniques to discover classification, clustering and outlier detection rules. In each of these tasks, we present the extracted knowledge and describe its importance in educational domain. Data mining techniques are analysis tools that can be used to extract meaningful knowledge from large data sets. This paper is designed to present and justify the capabilities of data mining in the context of higher educational system.

Keywords— Map-Reduce, Classification, Performanace, Data Mining, Analysis.

I. INTRODUCTION

Now a day's large quantities of data is being accumulated. Data mining is the process of discovering interesting knowledge from large amount of data stored in database or other information responsibility. The educational system in India is currently facing several issues such as identifying students need, personalization of training and predicting quality of student interactions. Educational data mining (EDM) provides a set of techniques which can help educational system to overcome this issue in order to improve Learning experience of students as well as increase their profits. Manual data analysis has been around for sometimes now, but it creates bottleneck for large data analysis. The transition won't occur automatically; in this case, we need the data mining. Data mining software allow user to analyzed data from different dimensions categorized it and summarized the relationship, identified during mining process .

The topic of explanation and Data analysis of academic performance is widely researched. Data Mining Techniques is the promising methodology to extract valuable information in this objective. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data. In this perspective, Data Mining can analyze relevant information results and produce different perspectives to understand more about the students' Performanace so as to customize the course for student learning.

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. In this perspective, one way to achieve highest level of quality in higher education system is discovering and analyzing relevant information results and produce different perspectives to understand more about the students' Performanace so as to customize the course for student learning.

Data mining task is used in Information Tegnology to facilitate students learning. Results are satisfactory because the existing technology aids and addresses the aspects of automated learning, practicing and evaluations of an academic cycle. They facilitate to understand/monitor students performance based on that internal and external examinition scores. At that time, there is no perfect usage of data mining techniques to facilitate Students Learning. So a better system is required to monitor and analyze student's performance based on a knowledge base constructed from automated learning, practicing and evaluations of the academic cycle.

Information's like intrenal and extrenal examinition scores were collected from the college ERP system, to measure the analysis performance at the end of the semester. This study will help the students and the teachers to improve the division of the student. This study will also work to identify Strong/weaker students and needed special attention to reduce fail percentages and taking appropriate action for the upcoming evaluations. This kind live performance monitoring and counter measures before the big evaluation definitely helps to improve students' performance.

In this paper, we used map-reduce based classification. This technique can give the valuable information for the decision-maker which is absent in tree based classifications.

II. RELATED WORK

Doctor and Iqbal [1], have proposed an Intelligent Framework for Monitoring Student Performance using Fuzzy Rule Based Linguistic Summarization. The fuzzy Linguistic Summarization (LS) technique is used to extract weighted

fuzzy rules from the clustered labeled data. Their system is used to evaluate new teaching approaches and methodologies more effectively by identify weaknesses to provide more personalized feedback on learner's progress. In addition, their system evaluates teaching factors effecting student performance, which is incorporated as part of an automated intelligent analysis and feedback system. However, their system has not considered students learning behaviors' by monitoring parameters on the learners' activity, emotion, personality traits with the performance to conduct extensive analysis of students learning behaviors'. Ryan S Baker [2], has used data mining technique to implement intelligent system. The availability of large datasets in usable formats emerged later in Education than other fields. School records were often stored in paper form in file cabinets up into the current millennium, and data from online learning systems was often in difficult-to-use formats, didn't store essential semantic information, or wasn't stored at all. Even today, a surprising amount of educational data is stored in the form of the screen locations of clicks and mouse movements, or perhaps with reference to arbitrarily named objects, making meaningful data analysis is difficult. However, storing educational data is not just whether a student was correct or incorrect, but the context of the behaviour and the skills involved in the task they were undertaking. Shargabi and Nusari[3], proposed Data Mining Techniques to discover vital patterns and calculate contribution of academic performance to help decision making. He has used Clustering (by K-means algorithm), association rules (by Apriori algorithm) and decision trees by (J48 and Id3 algorithms) techniques to build the data model. Quality assurance departments in higher education institutions could get a useful knowledge about the quality level of the different elements in education system using such techniques. Discovered patterns could be reformulated into academic strategies about many academic issues as admission policy, students' services and teaching methods.

Morais et al. [4] proposed an analytical approach to deal with e-learning data for "Monitoring Student Performance Using Data Clustering and Predictive Modeling". He has used Regression Methodology, K-means Algorithm and multivariate analysis technique. This system supports teacher decision-making process to understand profiles of answers in order to guide a student to future learning activities and identify which criteria, in each group. Validating the student's group in different context and investigate the subgroups where there was a high "Avg Assistance Score" at the beginning, but then was gradually lessened as the course moved along is not considered. In the dynamic perspective, to explore the way of clustering, for example cluster by the score of "Correct" or "Incorrect"

III. RESEARCH METHODOLOGY

Now-a-days, performance of individual student in educational system is evaluated based on the internal assessment and university examination. The internal assessment is calculated based on the performance of student in educational activities such as internals examination (IE), Labrotary work, and University examinition. The internal assessment is calculated by teachers. The university examination is one that is scored by the student in semester examination. To pass university examination, each and every student has to gain minimum marks both in internal as well as final examination in semester

3.1 Data Preparation:

The data set used in this study was obtained from B.E. students of BV's engineering college from session 2012 to 2016. The size of the data is 280 students records. In this step data stored in different files like examinition, Termwork, Practical and Internal Examinition.

Data collection is the first and foremost step in the Data Analysis process. To plan for quality education, it is necessary that any analysis navigate an ever-expanding sea of educational data and focus on gathering knowledge that can improve future prospects. Collected data must be preprocessed—that is, cleaned, transformed, and integrated before it can undergo the training process described next.

3.2 Methodology

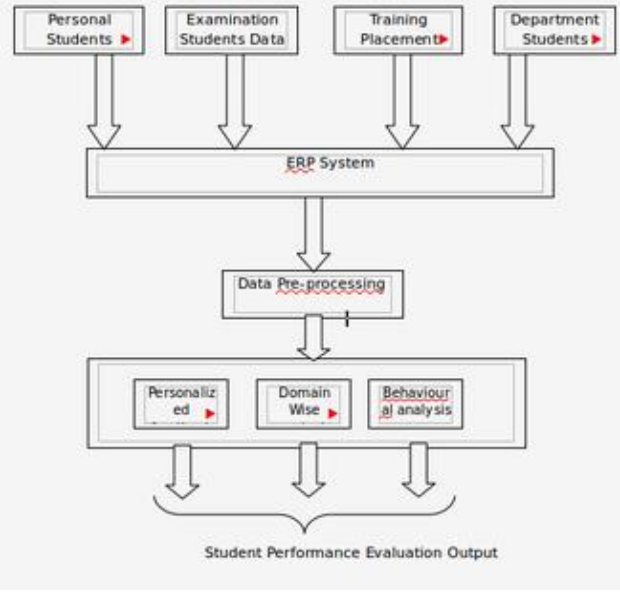

Figure1. Architecture of Student Performance Analysis System

This work proposes an extensional MapReduce Classification algorithm for performanance measure constraints. It allows user to specify a classification process in data mining to measure the indivisual and batch wise performanance measure and tries to make the data visualize for better understanding. This algorithm classifies the stuent data into several levels like Students Indivisual performanace, Strong/Weak Subject wise performanance and Domain wise student performanace. Under this algorithm, first it combines student data according to the input constaint of data level respectively. The experiments show that the student data classification algorithm can produce the correct student performanace. Secondly, it calculates the data mining process's average completion time which is based on the student data level. It improves the precision of classification's remaining time evaluation.

The Map Reduce classification algorithm takes the student data locality and cluster heterogeneity into account. The data locality is the key factor that affects the efficiency of MapReduce classification process. The data locality means that the classification's operation code and the classification's input data are on the same computing node or on the same rack. Of course, the efficiency when the code and data are on the same node is higher than on the same rack. If the code and data are on the same node, it would avoid the data transmission in the network and greatly reduce the delay. Therefore, in the large scale data processing applications, shifting the code would be "cheaper" than moving data. In this work, in order to meet the time constraints of the data mining process and further improve the data locality.

Following steps are carried out for measuring student performanance.

Step 1: Data gathering from BV's College of Engineering.

Step 2: Data preprocessing task is carried out.

Step 3: In MapReduce concept, Mapper partitions the student data with key value pair, after that Reducer aggregates the result . Consequently, the performance of MapReduce strongly depends on how it integrates the student data.

Step 4: Classification is done using map- reduce algoritham.

3.2.1 Data gathering from College

All the information used in this study has been gathered through ERP. In this research work academic dataset has been used. Its purpose was to obtain personalized, subject wise strong/weak and Domain wise information to identify some important factors that could affect college performance.

3.2.2 Data preprocessing

Before applying DM algorithm it is necessary to carry out some pre-processing tasks such as cleaning, integration, discretization and variable transformation. It must be pointed out that very important task in this work was data pre- processing, due to the quality and reliability of available information, which directly affects the results obtained. In fact, some specific pre-processing tasks were applied to prepare all the previously described data so that the classification task could be carried out correctly. First, all available student data .CSV files were laoded into the HDFS. After loading data perform data preprocessing steps for the indivisual, Subject wise weak/strong and domain wise student's information are analyzed. Furthermore, the continuous variables are transformed into discrete variables, which provide a much more comprehensible view of the data. In the following way dataset prepossed before mesuring the perfomances.

Subject_Total=Σsub[TH],sub[PR/OR],sub[TW],sub[IE]

Total_Marks_Subject[i]= $\sum_{i=0}^{n} subject[i][TH,PR/ORAL,TW]$

3.3 Classification using MapReduce

3.3.1 Basic MapReduce Terminology

In the proposed system the performance of the system is improved by using MapReduce. This is simple yet powerful framework which lets the programmer write simple units of work as map and reduce functions. In summary, they are:

- "In-mapper combining", where the functionality of the combiner is moved into the mapper. Instead of emitting intermediate output for every input key-value pair, the mapper aggregates partial results across multiple input records and only emits intermediate key-value pairs after some amount of local aggregation is performed.
- The related patterns "pairs" and "stripes" for keeping track of joint evens from a large number of observations. The pairs approach keeps track of each joint event separately, whereas the stripes approach keeps track of all events that co-occur with the same event. Although the stripes approach is significantly more efficient, it requires memory on the order of the size of the event space, which presents a scalability bottleneck.
- "Order inversion", where the main idea is to convert the sequencing of computations into a sorting problem. Through careful orchestration, one can send reducer the result of a computation (e.g., an aggregate statistic) before it encounters the data necessary to produce that computation.
- "Value-to-key conversion", which provides a scalable solution for secondary sorting. By moving part of the value into the key, one can exploit the MapReduce execution framework itself for sorting. Based on MapReduce concept, an extensional MapReduce Task Scheduling algorithm for performanance measure is proposed. It allows user to specify data mining process's for given constaints and tries to make the data mining process to be finished before the deadline.

3.3.2 Performance Measure using MapReduce

- **Individual Performance:** This performance was measured by applying following steps

1. Finding Subjects Code Semester Wise, Sub_Code[i]
2.Total Marks of Sub_Code[s]=Σ i=0 3 Sub_Codes[s];
3.Sem[i]=Total Marks of Sub_Codess[sub1 to sub5]
Where i=1 to 8;

- **Subject Wise Strong Performance:** This performance was measured by applying following steps

1. Data Preprocessing
2. Normalization
Threshold={[(Total_Marks_Subject[i]/Grand_Total_Subject[i])*60]/100};
if(Threshold>=60)Add(Strong_Vector[Subj_Code])
3.Sort_By_DESC (Strong_Vector[Subj_Code])
4. Display Strong_Vector[Subj_Code]

- **Subject Wise Week Performance:** This performance was measured by applying following steps

1. Data Preprocessing
2. Normalization
Threshold={[(Total_Marks_Subject[i]/Grand_Total_Subject[i])*60]/100};
if(Threshold<=40)Add(Strong_Vector[Subj_Code])
3.Sort_By_DESC(Strong_Vector[Subj_Code])
4. Display Strong_Vector[Subj_Code]

- **Domain Wise Performance:** This performance was measured by applying following steps

1. Data Preprocessing
2. Normalization
Domain Group_Creation
Domain_OS={OS,DS,COA}
Domain_PR={SPA,OOPM,WP,OST,AIT}
Domain_DB={DMS,ADBMS,DMBI,BDA}
Domain_NW={CN,SWS,CC,WT}
3. Total_OS= Σi=0 n OS[i],DS[i],COA[i]
Total_PR= Σi=0 n SPA[i],OOPM[i],WP[i],OST[i],AIT[i]
Total_DB= Σi=0 n DMS[i],ADBMS[i],DMBI[i],BDA[i]
Total_NW= Σi=0 n CN[i],SWA[i],CC[i],WT[i]
4. DPer_OS= (Total_OS/GT_OS)*100
DPer_PR= (Total_PR/GT_PR)*100
DPer_DB= (Total_DB/GT_DB)*100
DPer_NW= (Total_NW/GT_NW)*100
5. Total_Domain=Σ Dper_OS, Dper_PR, Dper_DB, Dper_NW
Dval_OS=(Domain_OS/ Total_Domain)*100;
Dval_PR=(Domain_PR/ Total_Domain)*100;
Dval_DB=(Domain_DB/ Total_Domain)*100;
Dval_NW=(Domain_NW/ Total_Domain)*100;

3.4 Attribute selection

The attributes are selected using the Feature Selection Techniques called, Correlation-based Feature Selection (CFS). The Correlation-based Feature Selection (CFS) estimates and ranks the subset of features than individual features. It chooses the set of attributes that are highly associated with the class, in addition to those attributes that are in low inter-correlation.

IV. RESULTS

The experiments shows that the student data classification map-reduce algorithm can improve the student data analysis process of the student performanace . The result shows that the proposed system has higher classification accuracy in big data and also reduces the time complexity.

Results and Discussion:

Here we analyze the performance offered classification with MapReduce based classification. The performance is evaluated by the parameters such as students indivisual achievement, subject wise strong/weak student performanace and domin wise student performanace. Based on the comparison and the results from the experiment, the proposed approach works better than the existing system.

The following graph shows the strong/weak and domain wise classification with MapReduce technique and proposed system such that MapReduce based classification based on two parameters of accuracy and methods such as existing and proposed system. From the graph we can see that, accuracy of the system is reduced somewhat in existing system than the proposed system. From this graph we can say that the accuracy of proposed system is increased which will be the best one

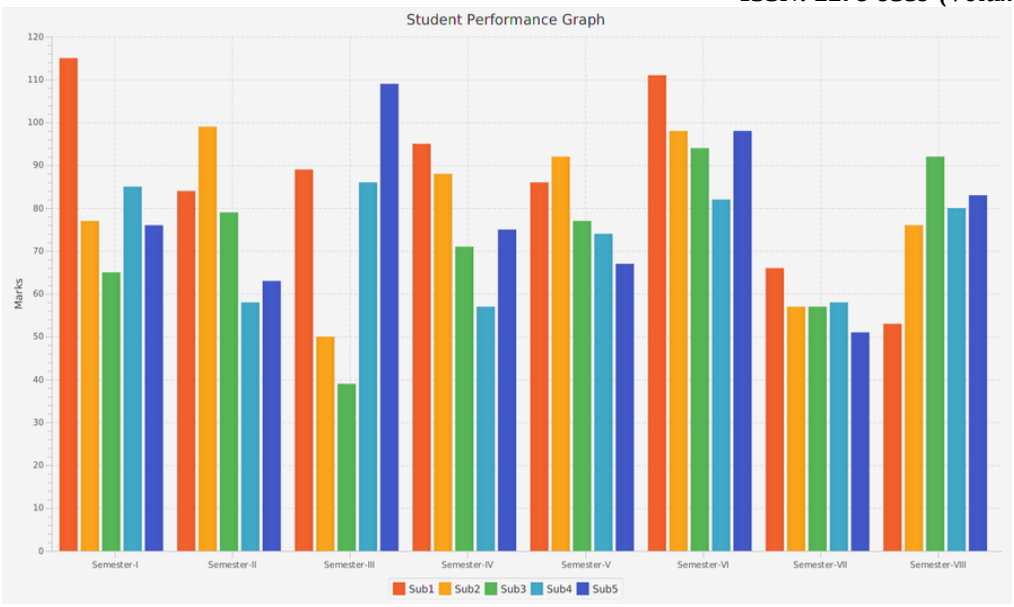

Figure 2. shows the Indivisual student performanace of sem I to VIII.

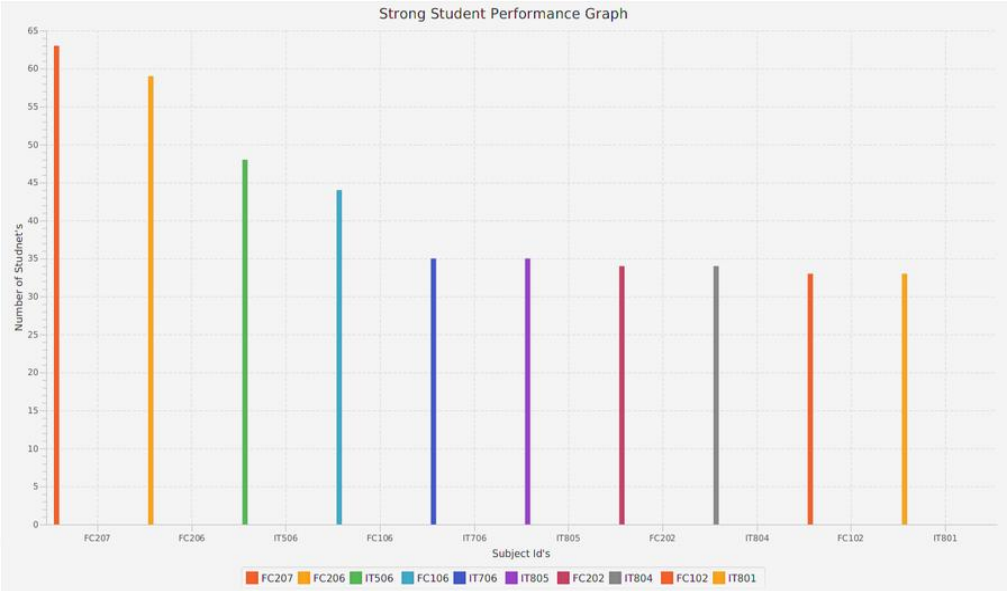

Figure 3. shows the subject wise strong student performanace.

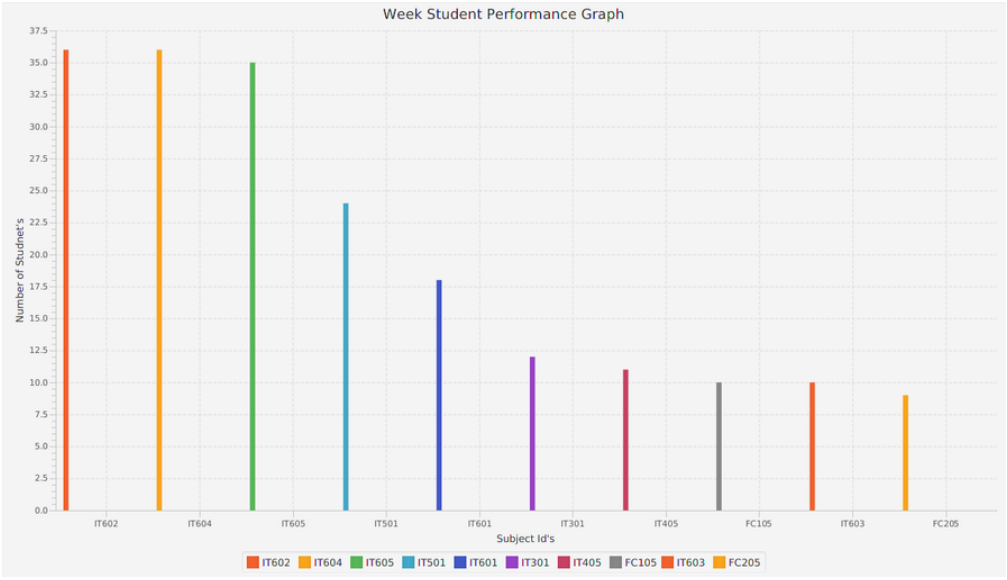

Figure 4. shows the subject wise weak student performanace.

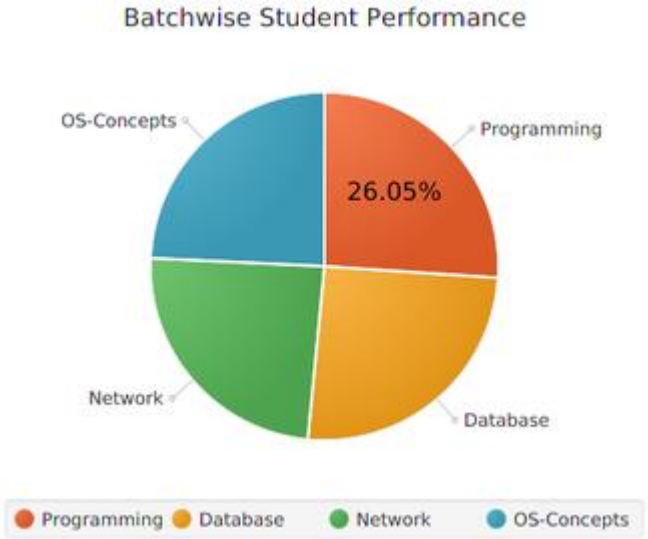

Figure 5. shows the Domain Wise strong and weak student performanace

V. CONCLUSIONS

The aim of this research study is to analyze the students academic performanace that contribute to the prediction of students' academic performance. It is useful in identifying weak students who are likely to perform poor in their studies. An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. The various data mining techniques can be effectively implemented on educational data. From the above results it is clear that classification techniques and mapreduce concept can be applied on educational data for predicting the student's outcome and improves their results. The classification accuracy and performance is high in the proposed system. This experiment shows that the proposed system is more efficient.

Finally, the next step in our research is to carry out experiments using more data and also from different educational levels (primary, secondary) to test whether the same performance results are obtained using different DM approaches.

ACKNOWLEDGMENT

I am thanful to my PhD Supervisor Professor, Dr. Praveen Kumar for helping me in preparing this paper and his constant support. I also acknowledge the help rendered to me by BVCOE Navi Mumbai for giving me permission to use student database of various examinitions.

REFERENCES

[1] Doctor, F. & Iqbal, R. "An intelligent framework for monitoring student performanceusing fuzzy rule-based Linguistic Summarisation Fuzzy Systems (FUZZ-IEEE)", 2012 IEEE International Conference on, 2012, 1-8.

[2] Baker, R. "Educational Data Mining: An Advance for Intelligent Systems in Education Intelligent Systems", IEEE, 2014, 29, 78-82.

[3] Al-shargabi, A. & Nusari, A. Discovering vital patterns from UST students data by applying data mining techniques Computer and Automation Engineering (ICCAE), 2010. The 2nd International Conference on, 2010, 2, 547-551.

[4] de Morais, A.; Araujo, J. & Costa, E. "Monitoring student performance using data clustering and predictive modelling Frontiers in Education Conference (FIE)", 2014 IEEE, 2014, 1-8.

[5] Piedade, M. & Santos, M. Business intelligence in higher education: Enhancing the teaching-learning process with a SRM system Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on, 2010, 1-5.

[6] Dass, R. Using Association Rule Mining for Behavioral Analysis of School Students: A Case from India System Sciences, 2009. HICSS '09. 42nd Hawaii InternationalConference on, 2009, 1-6.

[7] Ying-ting Zhu, FuZhang Wang, Xing-hua Shan &Xiao-yan Lv. K-mediods Clustering Based on MapReduce and Optimal Search of Mediods, 9th International Conference on Computer Science & Education (ICCSE), 2014,573-577.

[8] hua Zhu, X.; qiong Deng, Y. & ling Zeng, Q. The analysis on course grade of college-wide examination based on mixed weighted association rules mining algorithm Computer Application and System Modeling (ICCASM), 2010 International Conference on, 2010, 4, V14-530-V14-533.

[9] Spiess, J.; T'Joens, Y.; Dragnea, R.; Spencer, P. & Philippart, L. Using big data to improve customer experience and business performance Bell Labs Technical Journal, 2014, 18, 3-17.

[10] Chen, X.; Vorvoreanu, M. & Madhavan, K. Mining Social Media Data for Understanding Students Learning Experiences Learning Technologies, IEEE Transactions on, 2014, 246-259.

[11] Rodriguez Groba, A.; Vazquez Barreiros, B.; Lama, M.; Gewerc, A. & Mucientes, M. Using a learning analytics tool for evaluation in self-regulated learning Frontiers in Education Conference (FIE), 2014 IEEE, 2014, 1-8.

[12] Wu, X.; Zhu, X.; Wu, G.-Q. & Ding, W. Data mining with big data Knowledge and Data Engineering, IEEE Transactions on, 2014, 26, 97-107.
[13] Ahmadvand, A.; Bidgoli, B. & Akhondzadeh, E. A Hybrid Data Mining Model for Effective Citizen Relationship Management: A Case Study on Tehran Municipality e-Education, e-Business, e-Management, and e-Learning, 2010. IC4E '10. International Conference on, 2010, 277-281.
[14] Barros da Silva, H. & Leitao Adeodato, P. A data mining approach for preventing undergraduate students retention Neural Networks (IJCNN), The 2012 International Joint Conference on, 2012, 1-8.
[15] Knauf, R.; Sakurai, Y.; Takada, K. & Tsuruta, S. A Case Study On Using Personalized Data Mining For University Curricula Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, 2012, 3051-3056.
[16] hang, Y.; Chen, S.; Wang, Q. & Yu, G. i2MapReduce: Incremental MapReduce forMining Evolving Big Data Knowledge and Data Engineering, IEEE Transactions on, 2015, PP, 1-1.
[17] de C Gatti, M.; Vieira, M.; de Melo, J.; Cavalin, P. & Pinhanez, C. Handling big data on agent-based modeling of Online Social Networks with MapReduce Simulation Conference (WSC), 2014 Winter, 2014, 851-862.

