ISSN No. 0976-5697

Volume 2, No. 6, Nov-Dec 2011

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Educational Data Mining: An Emerging Trends in Education

Mrs Anita Chaware
Assistant Professor,
P.G.Department of Computer science,
SNDT Womens University,India
anitadongarwar@gmail.com

Abstract: Data mining over period of time is establishing itself as one of the major disciplines in computer science with growing industrial impact. Early Computing was to designing files to store the data so that information could be efficiently accessed but now the trend has completely changed. With enormous data in their pockets companies and organizations want to know about the patterns and trend that this data follows i.e. the knowledge to build the intelligence and take the action to improve their business. Data mining, the science of extracting useful knowledge from such huge data repositories, has emerged as a young and interdisciplinary field in computer science. Education researchers, with a large data repository in education field, are trying to make use of the concept of data mining. With this there is a rise in a field called Educational Data Mining (EDM).This research is to discover new knowledge or learner skills or features of education. This paper reviews the work done in this area.

Keywords: Higher Education, Education Data mining , Quality and Quantity of education, Tradition educational issues .

I. INTRODUCTION

Data mining, or the Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. Data mining is more than just conventional data analysis. Data mining, the intertwined of multiple disciplines, including statistics, machine learning, pattern recognition, database systems, information retrieval, World-Wide Web, Visualization,etc. [1]. Many applications areas such as banking, retail industry and marketing fraud detection, computer auditing, biomedical and DNA analysis, telecommunications, financial Industry, bioinformatics, and counter-terrorism.[2] have already been advanced through the sturdy techniques of data mining . It uses traditional analysis tools (like statistics and graphics) plus those associated with artificial intelligence (such as rule induction and neural nets).

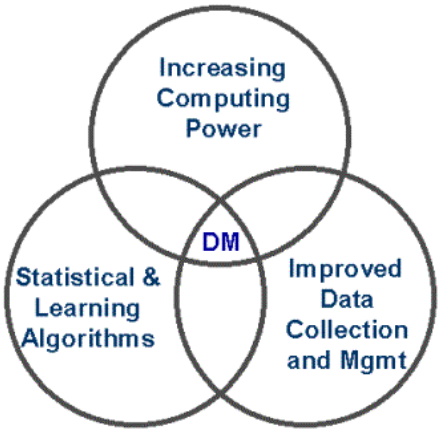

Figure: 1 Source : [35] DM→ Data Mining

The technologies behind data mining are the convergence of three basic concepts:

Increasing Computing Power
Improved data collection
Statistical and machine learning algorithms

With Powerful workstations and Cost effectiveness of servers are leading research to store and handle huge data repository with proper management. These huge data repository are the basis for applying different Statistical and machine learning algorithms to give the different data mining tools. Depending on the type of domain specific data, different Data mining tools are found to improve the domains feature or get a view of patterns this data is making which can be further used for decision making in this corporate world. One of the application of data mining on which this paper is focusing is education specifically the Higher Education.

This paper explains education data mining in Section 2. Section 3 explains about higher educational systems, its main objectives, the processes that happens regularly in there, and the main quality indicators in these systems. Then in section 4 the issues and methods of education data mining are explained. Section 5 is an analysis of the current works of data mining in education. In section 6, conclusion and further works are described.

II. EDUCATION DATA MINING

Over the past decade, new dynamics have emerged in each of the key domains of higher education, research and innovation which includes: (i) demand; (ii) diversification of provision; (iii) changing lifelong learning needs; and (iv) growing Information and Communication Technology (ICT) usage and enhanced networking and social engagement, both with the economic sector and with the community at large. [13].

With the use of Data mining algorithms it has become easier to study, analyze and predict issues including but not limited to enrollment management, dropout rate, estimated time to get degree, likes and dislikes of students while learning teaching process, student's behavior/performance, curriculum, usage of resources by individuals in education, identify excellent students for allocating scholarships and fellowships, identifying weak students who are likely to fail,

selection of best course for students, opening or closing of course, faculty improvement, human resource management. It suggests ways to improve courses and programs and attract more students in this ICT era and give them better Quality education.

III. ISSUES IN INDIAN HIGHER EDUCATION

Education in India is seen as one of the ways to upward social mobility. Good education is seen as a stepping stone to a high flying career. Education System in India currently represents a great paradox. One side due India's higher education policy of the 1950's, which envisaged schools of excellence, focusing more on technology and sciences, has finally paid off rich dividends. The creation of IITs, IIMs, NIT's, Schools of Science, Schools of Law, a large number of advanced training and research institutions like CDAC and NIT have globally been accepted [12] and on the other hand there are number of schools in the country that don't even have the basic infrastructure. With 60 plus years of independence we are far away from the goal of universal literacy. As per the United Nations Educational, Scientific and Cultural Organization [34] defines literacy as the "ability to identify, understand, interpret, create, communicate, compute and use printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, to develop their knowledge and potential, and to participate fully in their community and wider society." India has a billion-plus population and a high proportion of the young and hence it has a large formal education system.

The demand for education in developing countries like India has no limit as education is still regarded as an important bridge of social, economic and political mobility [9]. India being the third largest education system on the globe has to face the challenges to provide a quality of education to its own youth and attract other country youth as well as a part of globalization. The main issues in Indian Education is Quantity/Quality, along with the provision of education, higher and otherwise, to disadvantaged groups in Indian society (the issues of "inclusion" and "affirmative action"), Quality of education is a multi-dimensional concept, which should embrace all functions and activities, that is, teaching, academic programmes, research and scholarship, staffing, students, infrastructure, and academic environment. But India is lacking in nearly all of this fields of quality education. Barring a few premier institutions, the rest do not even have the capacity to meet the challenges of the new millennium.

Many research works are carried out in the improvement of education system abroad. This can be used in Indian universities where the issues are of Quality as well as Quantity also. With huge number of higher education aspirants, we believe that data mining technology can help bridging knowledge gap in higher educational systems. The hidden patterns, associations, and anomalies that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes.

IV. ISSUES AND METHODS IN EDUCATION DATA MINING

Educational data mining differs from knowledge discovery in other domains in several ways. This is because of the fact that due to very dynamic data available in educational field, it is difficult, or even impossible, to compare different methods or measures used in data mining and decide which is the best. The data varies a lot between samples, and teachers just cannot afford the time to access this data and do analyses on each sample, especially in real time. Therefore, as argued in [3], one should take care about the measures, parameters or methods used in educational data mining. Another difference is the size of the data: while tremendous amounts of data are collected about students' work, the size of the data on one sample is usually small. Typically in a classroom there are at best a few hundred students enrolled. Students may not all do the same exercises or activities. Collecting several years of data is certainly an option but there are instances where one wants to analyse the data as early as the first year. Besides, there are often changes between offerings of a course that have an impact on the common attributes of the data (for example not exactly the same topics/exercises/resources are offered from one year to the next). Therefore one should also be careful to avoid measures, parameters and methods where sample size has a predominant effect on the result.[3].

The methods used in Educational data mining are little different than normal data mining. Educational data mining methods are drawn from a variety of literatures, including data mining and machine learning, psychometrics and other areas of statistics, information visualization, and computational modeling.

As discussed in [4] these method can broadly classifies the work in educational data mining as given below:

- Prediction
- Classification
- Regression
- Density estimation
- Clustering
- Relationship mining
- Association rule mining
- Correlation mining
- Sequential pattern mining
- Causal data mining
- Distillation of data for human judgment
- Discovery with models

Source: Baker's taxonomy of educational data mining methods

The first three categories of Baker's taxonomy of educational data mining methods are same as the traditional data mining (the first set of sub-categories is directly drawn from Moore's categorization of data mining methods [5]). The fourth category, though not necessarily universally seen as data mining but more of , accords with Romero and Ventura's category of statistics and visualization, and has a prominent place both in published EDM research [3], and in theoretical discussions of educational data mining [4].

The fifth category of Baker's EDM taxonomy is perhaps the most unusual category, from a classical data mining perspective. In discovery with models, a model of a phenomenon is developed through any process that can be validated in some fashion (most commonly, prediction or knowledge engineering), and this model is then used as a component in another analysis, such as prediction or relationship mining. Discovery with models has become an increasingly popular method in EDM research, supporting sophisticated analyses such as which learning material sub-categories of students will most benefit from, [7] how different types of student behavior impact students' learning in different ways [8], and how variations in intelligent tutor design impact students' behavior over time .

V. ANALYSIS OF CURRENT WORK

There have been some studies done in the area of data mining in education. Each of them is trying to enhance the educational system by discovering patterns among the great deal of data. In this section, we analyze the existing works, the theme, the algorithms used, to get better understanding for our research work.

C.Marquez-vera,C. Romera and S. ventura[36] did a comparative study of ten different classification algorithms : five rule induction algorithms such as JRip, NNge, OneR, Prism and Ridor; and five decision tree algorithms such as J48, SimpleCart, ADTree, RandomTree and REPTree to predict the failure of school children who might fail and who might pass . The data collected was of school children and the results were compared from all algorithms to get the maximum output.

Waiyamai [15] has collected past 10 years data of students doing their major and minors, applied the association and classification rule of data mining technology to improve the quality of graduates and predict the most appropriate major for each single student. The main objectives in this survey is identifying whether **a** student is likely to be **a** good student in a given major or not, assisting in development of new curricula, and improving of existing curricula. The patterns found would be useful for the faculties of their institution to assist the new comer students while selecting their minors and majors.

Gabrielson [15] has used data mining to improve the score of students by analyzing the most effective factors in determining test score in various subjects. By applying two prediction algorithms, induction rule (Classification and Regression Tree) and neural network analyses, on the set of student record data of year 2002 matched with their test score, the patterns of most affected predictor variables is extracted. This patterns were used by the faculties to guide the students on their score improvement in coming year.

Jing Luan [17] used predicting and clustering the likelihood of persisters and non-persisters. The artificial neural network and rule induction (Classification and Regression Tree) **as** two prediction technique and Two-step as a clustering methods are used to apply on the set of student profile. The patterns found similar student through prediction technique, helped universities to facilitated in predicting the likelihood of student persistency. Data mining helps the universities to identify those students who are less likely to return to school year by year.. Therefore influence of data mining in marketing strategy increases the student's persistence rate.

Lan Wang1, Wei-Wei Zhuang-2, Yu-Fen Liu3[18] in their paper, based on the trend fitting model and the linear regression model, found the fact that the scale of graduate education in Hebel has greatly increased to a saturation status after the enrollment expansion in recent years and also predicts that the scale of graduate student enrollment should be compatible with the GDP growth rate. Also they suggested that postgraduate training mode must be reformed with innovation in the postgraduate education system, and that the capacity of high-level-talent training must be enhanced in accordance with the following ideas: "steady development, construction enhancement, structure optimization and quality guarantee", in which "steady development" has become the development keynote of the postgraduate education scale in Hebel Province.

Tong Yi in [19] had applied data mining software SPSS Clementine to two-step cluster analyzer to analyze teachers' position settlement of China's institution unit, selects 101 full-time teachers from School of Information Technology of JXUFE as research object, takes their teaching and research data in 2007 and 2009 into consideration and draw a conclusion that teachers are focused more on class hours and not scientific research. Higher Education teachers should emphasis more on research and for that they should be given points.

Emmanuel N. Ogor in his research paper[24] developed a methodology by the derivation of performance prediction indicators to deploying a simple student performance assessment and monitoring system within a teaching and learning environment by focusing on performance monitoring of students' continuous assessment (tests) and examination scores in order to predict their final achievement status upon graduation. The work used OLAP implementation with dynamic reporting capabilities and efficiency and also recommended for very large student databases in Oracle or MS SQL Server database environment to get accuracy.

In [25], Vasile Paul Bre_felean implemented J48 algorithm analysis tool on data collected from surveys on different specialization students ,with the purpose of differentiating and predicting their choice in continuing their education with post university studies (master degree, Ph.D. studies) through decision trees. He further concluded that clustering techniques, decisional trees for each specialization based on several algorithms, parallel analysis with the data collected could find the results that the courses having logical and analytical knowledge will be the first choice in future.

M. Vranic, D. Pintar and Z. Skocir in their paper[26] that the data gathering stage itself becomes one of the project and is a time consuming process because, gathering of various information about students, like details of their previous education, interests and other characteristic, collecting information such as this is not feasible with the organization support, also this data is stored in different format which needs to be processed before applying the data mining concepts . Finally the authors confirms that data mining and machine learning has a major role in education which need to be explored.

The paper[27], explores Visual Data Mining (VDM) model by combing visualization technology and data mining algorithm and apply it in the higher education evaluation system. The model they developed has the following

features, 1. Visual: It supports the visualization of data, mining process, and mining result. 2. Interactive: It makes full use of human's role played in the process of data mining. Users can select the areas they are interested in and analysis deeply. 3. Integrative: The model realized the integration of system and database, system and information manipulation and analysis, system and every mining task and algorithm, but still some issues are unaddressed like it can only deal with discrete data which is their future research work.

In [28] the authors did a survey of the specific application of data mining in learning management systems and a case study tutorial with the Moodle system. Also suggested that in e-learning data mining tools such as outlier analysis and social network analysis can be the most valuable. Also mining tools designed are application specific and difficult for end user to use it. Hence in future endues centric tools for education data mining should be design. Many more works are done in this area of e-learning and virtual campus using outliers and SNA.

In e-learning, outlier detection can be used for assisting instruction in the detection of learners' irregular learning processes[30], detecting a typical behavior in the grouping structure of the users of a virtual campus [31], detecting regularities and deviations in the learner's or educator's actions with others [32]

VI. FUTURE WORK AND CONCLUSIONS

Currently Indian education system is the third biggest education system in terms of enrollment, in the world. with billions of younger generation involved. Indian higher education issues are little different as compared to other developed countries. India has Quality as well as Quantity issues. With huge Quantity of student there is no problem of retention etc. but if we are unable to provide enough resource and Quality education and improve our educational Quality globally, soon this problem of retention and closing down institutions will come. Student will start approaching university outside India for their higher education, Graduation and Post Grad. With this data mining tools we can find some pattern and start taking decisions for improvements for the problems like the given below:

With giant industrial growth Is Indian youth getting a skill based education? Can we build some data mining tool to guide the students to select the courses.

The primary responsibility of the Country is to provide the eligible with good quality higher education at reasonable cost. This is in fact essential for the intellectual growth of the country. Can we predict or estimate the minimum cost a higher education aspirant should pay?

As per UGCs Guideline Appropriate and effective feedback mechanism (e.g. returning corrected answer books to students, responding to students' queries on the evaluation procedure, etc.) should be established at all institutions. Using these mechanisms, data mining tools can help institutions to work on the future improvement of course/curriculum best for student over a period of time.

We can measure the mental ability of students and predict which course the child should join so as to remove the educational stress among the students.

Using the overall funding patterns and trends, issue of institutional funding and student financing (student aid and loans) a model using data mining can be build to see the financial condition of the students (and capabilities of the institution while issuing the Government grants)and distribute the scholarships and grants to the deserving's.

Still in India ICT has not taken full fledge control and we are still in age of chalk and board, and other old traditional resources. A data mining model can be build to predict resources that will be needed in next coming 10 years and invest on them appropriately.

A model for predicting the best teacher of the year in advance depending on the data available, does the teacher needs any training depending on the feedback mechanism, the issues related to pay and leave of the teachers or higher education employee can be predicted and to take proper measure so as to retain the talent in this field.

There can be many more processes to consider and a model similar to [11] can be build looking at Indian education system problems which is our future work.

This model proposed by Naeimeh Delavaril Mohammad Reza Ayatollahzadeh Shirazi, Mohammad Reza Beikzadeh[10] can be used with variation to solve the issues in Indian Higher education system. Their model describe the main processes identified in higher educational systems: Each single process has some sub processes shown in the second column, followed by the knowledge which we will gather from the educational data (students, academic, faculties and administrative). The fourth column of the table focuses on the enhancement of the traditional educational processes. The last column of the table portrays the most appropriate data mining techniques in order for extracting the beneficial knowledge. For example, "evaluation" is an educational process. Its main sub-processes are "student assessment", "lecturer assessment", "industrial training assessment", "course assessment", and "student's registration evaluation". The prediction techniques such as linear and multiple regression or neural network can be applied to the set of student data in student assessment sub-process; and the success patterns of previous similar students can be extracted. The result could be able the universities / institution to predict the likelihood of student's persistence.

Finally to conclude, this paper has been an effort in providing the motivation toward advancement of the traditional educational process via data mining technology. The main idea of this analysis was a survey which targets the superior advantages of data mining in each single higher educational system process. It also provides an opportunity for us to be known with the existing area of study for data mining in education. In India, higher education is in Quest of Quality locally as well as globally, but in view of lack of resources to meet the increasing demand, there's a need of solving this problems with the data mining techniques, as quantity of data is not the issue.

VII. REFERENCES

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann, 2006.

[2] http://www.ercim.eu/publication/ws-proceedings/12th-EDRG/ EDRG12_Re.pdf

[3] Beck J. Difficulties in inferring student knowledge from observations (and why you should care). Educational Data Mining workshop, in conjunction with 13th International

Conference of Artificial Intelligence in Education, Marina del Rey, CA. USA. July 2007, pp.21-30.

[4] Baker, R.S.J.d., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1 (1), 3-17.

[5] Beck, J. and Woolf, B. 2000. High-level student modeling with machine learning. In Proceedings of the International Conference on Intelligent tutoring systems, 584-593.

[6] Beck J.E. 2007. Difficulties in inferring student knowledge from observations (and why you should care). Proceedings of the AIED2007 Workshop on Educational Data Mining, 21-30.

[7] Beck, J.E. and Mostow, J. 2008. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 353-362.

[8] M. Cocea and S. Weibelzahl, "Can log files analysis estimate learners' level of motivation?" in Proc. Workshop week Lernen— Wissensent deckung- Adaptivit ¨at, Hildesheim, Germany, 06, pp. 32–35.

[9] Amutabi, M. N. & Oketch, M. O. (2003), 'Experimenting in distance education: the African Virtual University (AVU) and the paradox of the World Bank in Kenya', International Journal of Educational Development **23**(1),

[10] Naeimeh Delavaril, Mohammad Reza Ayatollahzadeh Shirazi, Mohammad Reza Beikzadeh," **A** New Model for Using Data Mining Technology in Higher Educational Systems" Information Technology Based Higher Education and Training, 2004. ITHET 2004.

[11] Delavari N. & Beikzadeh M. R & Shirazi M. R. A., "A New Model for Using Data Mining in Higher Educational System", in Proceedings of 5th International Conference on Information Technology Based Higher Education and Training (ITHET), Istanbul, Turkey, May 31 to June 2, 2004.

[12] Sanat Kaul," Higher Education In India: Seizing The Opportunity", Indian Council For Research On International Economic Relations, May 2006

[13] V. Lynn Meek,Ulrich Teichler,Mary-Louise Kearney "Higher Education, Research and Innovation: Changing Dynamics", Report on the UNESCO Forum on Higher Education, Research and Knowledge 2001-2009

[14] C. Marquez-Vera, C. Romero and S. Ventura ,"Predicting School Failure Using Data Mining", http://educationaldatamining.org /EDM2011/wp-content/ uploads/proc

[15] K. Waiyamai. "Improving Quality **of** Graduate Smdents by Data Mining".Dept. **of** Computer Engineering, Faculty of Engineering, KasetsaR Univcrsity, Bangkok, Thailand, 2003

[16] **S.** Gabrilson, "Dah Mining with CRCT Scorer". Office of inforrna. Technology, Geogia Department of Education. October 15,2003.

[17] J.,Luan, "Data Mining and Knowledge Management in Higher Education- Potential Applications",Proceedings of AIR Forum, Toronto, Canada, 2002.

[18] Lan Wang1, Wei-Wei Zhuang-2, Yu-Fen Liu3, " An Emperical study on the prediction model of Postgraduate education in Hebel province" , Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

[19] Tong Yi, "Apply Data Mining to Predict Teachers' Position Settlement: A Case Study of University", 978-1-4244-5540-9/10/$26.00 ©2010 IEEE

[20] Muslihah Wook, Yuhanim Hani Yahaya, Norshahriah Wahab, Mohd Rizal Mohd Isa, Nor Fatimah Awang, Hoo Yann Seong, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques", 2009 Second International Conference on Computer and Electrical Engineering, 978-0-7695-3925-6/09 $26.00 © 2009 IEEE

[21] Ning Fang and Jingui Lu, "Work in Progress - A Decision Tree Approach to Predicting Student Performance in A High-Enrollment, High-Impact, and Core Engineering Course", 39th ASEE/IEEE Frontiers in Education Conference, 978-1-4244-4714-5/09/$25.00 ©2009 IEEE

[22] Chang-xin Song, Ke Ma, "Applications of Data Mining in the Education Resource Based on XML", 2008 International Conference on Advanced Computer Theory and Engineering, 978-0-7695-3489-3/08 $25.00 © 2008 IEEE

[23] Song Lihua, Zhao Yongsheng, Zhang Zhonglei, "Research on data mining in college education", 2008 International Conference on Computer Science and Software Engineering,978-0-7695-3336-0/08 $25.00 © 2008 IEEE

[24] Emmanuel N. Ogor, "Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques", Fourth Congress of Electronics, Robotics and Automotive Mechanics, 2007, IEEE

[25] Vasile Paul Bre_felean , "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment", Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28, 2007, Cavtat, Croatia

[26] M. Vranic, D. Pintar and Z. Skocir, "The use of data mining in education environment ",9th International Conference on Telecommunications - ConTEL 2007

[27] Hanjun Jin, Tianzhen Wu, Zhiliang Liu and ianlin Yan, "Application of Visual Data Mining in Higher-education Evaluation System", 2009 First International Workshop on Education Technology and Computer Science.

[28] Cristóbal Romero , Sebastián Ventura, Enrique García , " Data mining in course management systems: Moodle case study and tutorial", Submitted to Elsevier Science

[29] Ueno, M., (2004a). Data mining and text mining technologies for collaborative learning in an ilms "samurai". In IEEE International Conference on Advanced Learning Technologies, Joensuu, Finland (pp. 1052-1053).

[30] Ueno, M., (2004b). "Online outlier detection system for learning time data in e-learning and its evaluation", In International Conference on Computers and Advanced Technology in Education, Beijing, China (pp. 248–253).

[31] Castro, F., Vellido, A., Nebot, A., Minguillon, J., 2005. "Detecting atypical student behaviour on an e-learning system", In Simposio Nacional de Tecnologas de la Informacin y las Comunicaciones en la Educacion, Granada, Spain (pp. 153–160).

[32] Muehlenbrock, M. (2005). "Automatic action analysis in an interactive learning environment", In Proceedings of the workshop on Usage Analysis in Learning Systems at the 12th International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands (pp. 73-80).

[33] Scott, J. (2000). Social Network Analysis: A Handbook 2nd Ed. Newberry Park, CA: Sage.

[34] http://www.unesco.org/new/en/education/themes/education-building-blocks/literacy/

[35] http://inspa.info/pdf/Indian%20Edu%20Sysytem.pdf

[36] http://www.thearling.com/dmintro/dmintro_2.htm,