A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS

Rajni Jindal and Malaya Dutta Borah

Department of Computer Engineering, Delhi Technological University, N. Delhi, India
rajnijindal@dce.ac.in
malayadutta@dce.ac.in

ABSTRACT
Educational Data Mining (EDM) is an emerging field exploring data in educational context by applying different Data Mining (DM) techniques/tools. It provides intrinsic knowledge of teaching and learning process for effective education planning. In this survey work focuses on components, research trends (1998 to 2012) of EDM highlighting its related Tools, Techniques and educational Outcomes. It also highlights the Challenges EDM.

KEYWORDS
Educational Data Mining (EDM), EDM Components, DM Methods, Education Planning

1. INTRODUCTION
Educational Data Mining (EDM) is an emerging field exploring data in educational context by applying different Data Mining (DM) techniques/tools. EDM inherits properties from areas like Learning Analytics, Psychometrics, Artificial Intelligence, Information Technology, Machine learning, Statics, Database Management System, Computing and Data Mining. It can be considered as interdisciplinary research field which provides intrinsic knowledge of teaching and learning process for effective education [18].

The exponential growth of educational data [37] from heterogeneous sources results an urgent need for research in EDM. This can help to meet the objectives and to determine specific goals of education. EDM objective can be classified in the following way:

(1) Academic Objectives
—Person oriented (related to direct participation in teaching and learning process)

E.g.: Student learning, cognitive learning, modelling, behavior, risk, performance analysis, predicting right enrollment decision etc. both in traditional and digital environment and Faculty modelling- job performance and satisfaction analysis.

—Department/Institutions oriented (related to particular department/institutions with respect to time, sequence and demand).

E.g.: Redesign new courses according to industry requirements, identify realistic problems to effective research and learning process.

—Domain Oriented (related to a particular branch/institutions)
E.g.: Designing Methods-Tools, Techniques, Knowledge Discovery based Decision Support System (KDDS) for specific application, branch and institutions.

(2) Administrative Objectives

—Administrator Oriented (related to direct involvement of higher authorities/administrator)

E.g.: Resource (Infrastructure as well as Human) utilization, Industry academia relationship, marketing for student enrollment in case of private institutions and establishment of network for innovative research and practices.

—To explore heterogeneous educational data by analyzing the authors’ views from traditional to intelligent educational systems in the decision making process.

—To explore intelligent tools and techniques used in EDM and

—To find out the various EDM challenges.

To meet academic and administrative objectives, a survey of EDM is necessary which focus on cutting edge technologies for quality education delivery. This paper discusses the EDM components and research trends of DM in Educational System for the year 1998 to 2012 covering various issues and challenges on EDM.

The rest of this paper is organized into 5 sections. Section-2 focuses on EDM Components such as Stakeholders, environments, data, methods, tools etc. Section-3 is about mining educational objectives. Section-4 highlights the research trends in EDM including various authors’ views in educational outcomes, useful EDM Tools and Techniques. Section-5 is a discussion based on section 3 and 4, Section-6 concludes the paper with observations based on the survey work and the future scope of EDM.

2. EDM COMPONENTS

The key components of EDM are Stakeholders of Education, DM Methods-Tools and Techniques, Educational data, Educational task and Outcomes which meet the Educational objectives (see fig. 1).

![Figure1. EDM components](image)

2.1. Stakeholders

Considering primary to higher education, major stakeholders of education can be divided in three groups:
Primary group. This group is directly involved with teaching and learning process. E.g.: Students (learners) and Faculties (teachers / learners, educators etc.)

Secondary group. This group is indirectly involved in the growth of the institution. E.g.: Parents and Alumni.

Hybrid group. This group is involved with administrative/decision making process e.g.: Employers, Administrator/Educational Planner, and Experts.

2.2. EDM environments

Formal Environment. Direct interaction with primary group stakeholder of education. E.g.: face to face classroom interaction.

Informal Environment. Indirect interaction with primary group stakeholder of education. E.g.: web based education (e-learning [14] e-training used in Chu et al. [32], online supported tasks)

Computer Supported Environment (individual and interaction). Direct and /or Indirect interaction with all the three groups (depends upon the objectives) stakeholder of education. E.g.: Intelligent Tutoring Systems- Tools such as DOCENT, IDE, ISD Expert, Expert CML related to curriculum development [67].

Tools such as such as Algebra Tutor, Mathematics Tutor, eTeacher, ZOSMAT , REALP, CIRCSIM-Tutor, Why2-Atlas, SmartTutor, AutoTutor, ActiveMath, Eon, GTE, REDEEM related to tutoring system.

—Collaborative learning used in [55]
—Adaptive Educational System [1]

2.3. Educational Data

Decision-making in the field of academic planning involves extensive analysis of huge volumes of educational data [63]. Data’s are generated from heterogeneous sources like diverse and varied uses in [44], diverse and distributed, structured and unstructured data.

These data’s are mostly generated from the offline or online source

Offline Data. Offline Data are generated from traditional and modern classroom interaction interactive teaching/learning environments, learner/educators information, students attendance, emotional data, course information, data collected from the academic section of an institution [18] etc..

Online Data. Online Data are generated from the geographically separated stake holder of the education, distance educations used in [15], web based education used in [31], computer-supported collaborative learning used in [55] social networking sites and online group forum.

E.g: Web logs, E-mail, Spreadsheets, Transcribed Telephonic Conversations, Medical records, Legal Information, Corporate contracts, Text data, publication databases[69] etc.
Uncertain Data. Uncertain Data are generated from scientific measurement techniques and heterogeneity in designing Data Warehouse (DWH) [23], sensor generated data, privacy preservation process data, summarization of data [10].

2.4. Educational Task

It is a continual process for formation of Vision and Mission of an institution, to nurture the talent of students which addresses issues in a responsive, ethical and innovative manner to meet the academic and administrative objectives. This task can divide into two types:

Decision making task. Active participation of the hybrid group of stakeholder to fulfill administrative oriented objectives.

Learner based task. Active participation of Primary stakeholder to fulfill academic objectives.

2.5. DM Methods

DM methods are one of the main components in EDM. As per the different purpose it can be broadly divided into two groups [53]:

— Verification Oriented (Traditional Statistics- Hypothesis test, Goodness of fit, Analysis of Variance etc.)

— Discovery Oriented (Prediction and Description- Classification, Clustering, Prediction, Relationship Mining, Neural Network, Web mining etc.)

Following DM methods are popular with the EDM research community.

Classification

It is a two way technique (training and testing) which maps data into a predefined class. This technique is useful for success analysis with low, medium, high risk students used in [37], student monitoring systems [42], predicting student performance, misuse detection used in [6] etc.

Statistics

It is a technique to identify outlier fields, record using mean, mode etc. and hypothetical testing. This technique is useful to improve the course management system & student response system [36].

Clustering

It is a technique to group similar data into clusters in a way that groups are not predefined. This technique is useful to distinguish learner with their preference in using interactive multimedia system used in [12], Students comprehensive character analysis used in [75] and suitable for collaborative learning used in [24,7].

Prediction

It is a technique which predicts a future state rather than a current state. This technique is useful to predict success rate, drop out used in Dekker et al. [30,75], and retention management used in [61] of students.
Neural Network

It is a technique to improve the interpretability of the learned network by using extracted rules for learning networks. This technique is useful to determine residency, ethnicity used in [71], to predict academic performance used in [37], accuracy prediction in the branch selection used in [45] and explores learning performance in a TESL based e-learning system [68].

Association Rule Mining

It is a technique to identify specific relationships among data. This technique is useful to identify students’ failure patterns [52], parameters related to the admission process, migration, contribution of alumni, student assessment, co-relation between different group of students, to guide a search for a better fitting transfer model of student learning etc. used in [26].

Web mining

It is a technique for mining web data. This technique is useful for building virtual community in computational Intelligence used in [38], to determine misconception of learners used in [29] and to explore cognitive sense.

Apart from the above methods, [7] mentioned two new methods i.e. distillation of data for human judgment and discovery with models to analyze the behavioral impact of students in learning environments [14].

2.6. Tools

Due to the rapid growth of educational data, there is a need to summarize the tools according to their function/features, integrated techniques and working platforms. EPRules, GISMO, TADA-ED, O3R, Synergy/ColAT, LISTEN Mining tool, MINEL, LOCO, CIECoF, PDinamet, Meerkat, MMT tool are examples of EDM tools [18]. Apart from these, this research trying to find out more tools which will helpful for the research community to mine educational data as per requirements (see Table-1)

<table>
<thead>
<tr>
<th>Name of Tool and Developer</th>
<th>Source (Open/free/Commercial)</th>
<th>Function/Features</th>
<th>Techniques/Tools</th>
<th>Environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligent Miner (IBM)</td>
<td>Commercial</td>
<td>Provides tight integration with IBB’s DB2 relational db system, Scalability of Mining Algorithm</td>
<td>Association Mining, Classification, Regression, Predictive Modelling, Deviation detection, Clustering, Sequential Pattern Analysis</td>
<td>Windows, Solaris, Linux</td>
</tr>
<tr>
<td>MSSQL Server 2005 (Microsoft)</td>
<td>Commercial</td>
<td>Provides DM functions both in relational db system and Data Warehouse (DWH) system environment.</td>
<td>Integrates the algorithms developed by third party vendors and application users.</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>Tool Name</td>
<td>Type</td>
<td>Features</td>
<td>Analysis Capabilities</td>
<td>Operating Systems</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>MineSet (SGI)</td>
<td>Commercial</td>
<td>Provides Robust Graphics tools such as rule visualize, Tree visualize, Map visualize and scatter visualize</td>
<td>Association Mining, Classification, advanced statistics and visualization tools</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>Oracle Data Mining (Oracle Corporation)</td>
<td>Commercial</td>
<td>Provides an embedded DWH infrastructure for multidimensional data analysis</td>
<td>Association Mining, Classification, Prediction, Regression, Clustering, Sequence similarity search and analysis.</td>
<td>Windows, Mac, Linux</td>
</tr>
<tr>
<td>SPSS Clementine (IBM)</td>
<td>Commercial</td>
<td>Provides an integrated data mining development environment for end users and developers.</td>
<td>Association Mining, Clustering, Classification, Prediction and visualization tools</td>
<td>Windows, Solaris, Linux</td>
</tr>
<tr>
<td>Enterprise Miner (SAS Institute)</td>
<td>Commercial</td>
<td>Provides variety of statistical analysis tools</td>
<td>Association Mining, Classification, Regression, Time-series analysis, Statistical analysis, Clustering</td>
<td>Windows, Solaris, Linux</td>
</tr>
<tr>
<td>Insightful Miner (Insightful Incorporation)</td>
<td>Commercial</td>
<td>Provides visual interface, which allows users to wire components together to create self-documenting programs</td>
<td>Data cleaning, Clustering, Classification, Prediction, Statistical analysis</td>
<td>Windows, Solaris, Linux</td>
</tr>
<tr>
<td>CART (Salford Systems)</td>
<td>Commercial</td>
<td>Provides binary splitting and post pruning for Classification (Decision Tree) and for Prediction (Regression Trees).</td>
<td>Classification-Decision and Regression Tree</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>TreeNet(R) (Salford Systems)</td>
<td>Commercial</td>
<td>Provides an automatic selection of candidate predictors, Ability to handle data without preprocessing.</td>
<td>Classification, regression</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>RandomForests (Salford Systems)</td>
<td>Commercial</td>
<td>Provides high levels of predictive accuracy and an innovative set of graphical displays to reveal unexpected patterns in data</td>
<td>Clustering</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>GeneSight (Inc. of El Segundo, CA)</td>
<td>Commercial</td>
<td>Provides the researcher to explore large data sets from multiple experimental groups using advanced normalization, visualization, and statistical decision support tools</td>
<td>Visualization, K-means, and Neural Network Clustering, Time series analysis</td>
<td>Windows, Mac, Linux</td>
</tr>
<tr>
<td>Software Name</td>
<td>License</td>
<td>Description</td>
<td>Features</td>
<td>Platform</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>PolyAnalyst</td>
<td>Commercial</td>
<td>Provides decision maker to derive knowledge from large volumes of text and structured data</td>
<td>Clustering, Classification, Prediction, Association Mining</td>
<td>Windows</td>
</tr>
<tr>
<td>iData Analyzer (Microsoft)</td>
<td>Open /Free</td>
<td>Provides platform for visual learning environment</td>
<td>Pre processor, ESX, Heuristic Agent, Neural Network, Rule Maker and Report generator</td>
<td>Windows, Linux, Solaris</td>
</tr>
<tr>
<td>See5 and C5.0 (RuleQuest)</td>
<td>Open/Free</td>
<td>Provides Decision Tree Analysis, Commercial version of C4.5 DT algorithm</td>
<td>Decision Tree</td>
<td>Windows , Unix</td>
</tr>
<tr>
<td>TANAGRA (SPAD)</td>
<td>Open/Free</td>
<td>Provides data analyses in the early 90s Free Data Mining software for academic purpose. Ability to design of the GUI, Adding new algorithm</td>
<td>Factor analysis of mixed data, Support Vector Machines algorithms, Non iterative Principal Factor Analysis</td>
<td>Windows, Linux, Mac OS, Solaris</td>
</tr>
<tr>
<td>SIPINA (Ricco Rakotomalala Lyon, France)</td>
<td>Open/Free</td>
<td>Provides an environment for supervised learning algorithms, handle both continues &amp; discrete data</td>
<td>DT algorithms - C4.5,ID3</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>ORANGE (University of Ljubljana, Slovenia)</td>
<td>Open/Free</td>
<td>Provides open source data visualization and analysis for novice and experts</td>
<td>Text mining and Bioinformatics add-ons</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>ALPHA MINER (E-Business Technology Institute)</td>
<td>Open/Free</td>
<td>Provides the best cost-and-performance ratio for data mining applications</td>
<td>Versatile data mining functions</td>
<td>Windows ,Linux, Mac</td>
</tr>
<tr>
<td>WEKA (University of Waikato, New Zealand)</td>
<td>Open/Free</td>
<td>Provides machine learning algorithms for data mining tasks. Well-suited for developing new machine learning schemes.</td>
<td>Data pre-processing, classification, regression, clustering, association rules, and visualization.</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>Carrot</td>
<td>Open/Free</td>
<td>Provides ready-to-use components for fetching search results from various sources</td>
<td>Clustering</td>
<td>Windows, Linux</td>
</tr>
</tbody>
</table>

### 2.7. Data Links

One of the academic objectives of EDM is domain specific. This study tries to find out some suitable dataset/links as suitable for different domains of EDM (see Table-2).
### Table 2. Useful Dataset/Links

<table>
<thead>
<tr>
<th>Name</th>
<th>Web Address</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSLC Data shop</td>
<td><a href="http://pslcdatashop.org">http://pslcdatashop.org</a></td>
<td>Educational Data Mining. Collected data from different online learning environments.</td>
</tr>
<tr>
<td>JSE Data Archive</td>
<td><a href="http://www.amstst.org/publications/jse/toc.html">http://www.amstst.org/publications/jse/toc.html</a></td>
<td>Contains data archive for different DM applications</td>
</tr>
<tr>
<td>KD Nuggets</td>
<td><a href="http://kdnuggets.com">http://kdnuggets.com</a></td>
<td>Data Mining and Knowledge Discovery</td>
</tr>
<tr>
<td>DASL</td>
<td><a href="http://lib.stat.cmu.edu/DASL">http://lib.stat.cmu.edu/DASL</a></td>
<td>Illustrates Statistical methods</td>
</tr>
<tr>
<td>ML Net Ois</td>
<td><a href="http://www.minet.org">http://www.minet.org</a></td>
<td>Data Mining and Knowledge Discovery</td>
</tr>
<tr>
<td>UCI Machine Learning Repository</td>
<td><a href="http://www.ics.uci.edu">http://www.ics.uci.edu</a></td>
<td>iDA Dataset package for Higher Education such as:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>i. Medical-(CardiologyCategorical.xls, CardiologyNumerical.xls, SpineData.xls)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ii. Wildlife Management -(DeerHuntewr.xls)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>iii. Astronomy-(grb4u.xls)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>iv. Finance-(NasdaqDow.xls)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>v. Geography-(sonar.xls,sonaru.xls, UsTemperatures.xls)</td>
</tr>
<tr>
<td>Visualization and Data Exploration</td>
<td><a href="http://www-958.ibm.com/software/data/cognos">http://www-958.ibm.com/software/data/cognos</a> / manyeyes/</td>
<td>Explore existing visualized datasets and upload their own for exploration.</td>
</tr>
<tr>
<td></td>
<td><a href="http://hint.fm/">http://hint.fm/</a></td>
<td>Data visualization</td>
</tr>
<tr>
<td></td>
<td><a href="http://research.uow.edu.au/lea">http://research.uow.edu.au/lea</a> rningnetworks/seeing/snapp/index.html</td>
<td>Visualizing networks resulting from the posts and replies to the discussion forums</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.socialexplorer.co">http://www.socialexplorer.co</a> m/</td>
<td>Visualizations of census data and demographic information</td>
</tr>
<tr>
<td>Online Learning Systems with Analytics</td>
<td><a href="http://www.assistments.org">http://www.assistments.org</a>.</td>
<td>Helps teachers write assessments and then see reports on how their students performed</td>
</tr>
<tr>
<td></td>
<td><a href="http://wayangoutpost.com/">http://wayangoutpost.com/</a></td>
<td>Intelligent Tutoring System</td>
</tr>
<tr>
<td></td>
<td><a href="http://oli.web.cmu.edu/openlearn">http://oli.web.cmu.edu/openlearn</a> ing/forstudents/freecourses .</td>
<td>Offers open and free courses</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.khanacademy.org">http://www.khanacademy.org</a></td>
<td>Provides a platform for educational stakeholders</td>
</tr>
</tbody>
</table>

### 3. MINING EDUCATIONAL OBJECTIVES

This survey focused on mining academic objectives of EDM in context of traditional to dynamic environments.

In traditional teaching and learning environment Performance and Behavior analysis are performed on the basis of observation and paper records used in .[15,60]. This process is static used in [44]. This system has the drawbacks such as it cannot meet the need of the individual learner as well as lacking dynamic learning which can be improved by using five steps of an academic analytics process such as capture, report, predict, act and refine [36].

Learning and assessment process in a virtual environment using sophisticated DM methods in a digital learning environment is presented by [8]. This research focused on individual learners by “information-processing narratives” and group learners by “socio-cognitive narratives”.

To enhance the quality in higher learning institutions, the concept of predictive and descriptive models discussed by [51]. Predictive model predicts the success rate for individual students;
individual lecturer and Descriptive model describe the pattern modeling of student course enrollment, course assignment policy making, behavior analysis etc.

In Web-based education system learner behaviors, access patterns are recorded in a log file described in [62], hence able to analyze the need of the individual learner. To better design and modification of web sites by analyzing the access patterns in weblogs are described in [33].

The limitation of the log file is the authenticity of the user.

In [71] provides the different way to log record process by keeping record of learning path. This approach is suitable only for small log files.

To accumulate large log file in a real or virtual environment, an approach given by [50] where recording of all activities of learning such as reading, writing, appearing test, communicate with peer groups are possible.

To enhance this concept, [43] added collaborative learning approach between learner groups and educators which provides an easy way to analysis learner learning behavior.

E-learning is one way of mining online data. Importance of DM in e-learning, concept map in e-learning described in [11] learning management and Moodle system was described in [16].

Researchers [43,34,70] consider the “perception behavior” of learners and analysis with the help of sequential pattern mining technique which is able to analysis the data in a time sequence of actions.

The researchers mixed up the different DM techniques to validate the Predictive and Descriptive model so it is not clearly visible which technique/algorithm is to discover the appropriate quality in higher education.

To overcome this issue, an approach given by [4], trying to discover the vital patterns of students by analyzing academic and financial data in terms of validity, reality, utility and originality. The researcher used clustering algorithm (k-means), Association Rules (Apriori algorithm) and DT algorithm (J48, ID3) and WEKA data mining tool to validate the data model. In this research, researcher focuses on vital pattern analysis in higher education system. But researcher did not mention which algorithm/technique is best to analyze vital patterns for quality education.

Knowledge based decision technique by comparative study of the DM algorithms (C5.0 CART, ANN) and DM Tool (SPSS Clementine) was given by [20]. Attribute mainly considered in this research work was enrollment decision making parameters such as parental pressure, demand of industry and historical placement record. To enhance the accuracy of the analysis real data set of AIEEE 2007 was used in this work. This research work concluded that C5.0 has the highest accuracy rate to predict the enrollment decision. Another approach given by [45], proposing a new Attribute Selection Measure Function (heuristic) on existing C4.5 algorithm. The advantage of heuristic is that the split information never approaches zero, hence produces stable Rule Set and Decision Tree.

Most of the above discussed researches try to meet student perspectives, where as analyses of satisfaction levels of teachers were not discussed which is also important in the educational system.

To analyze this matter [40] proposed a model which comprises of five attributes i.e. Positive affect, Goal support, Self efficacy, Work conditions and Goal progress.
This model tested on the sample data of Abu Dhabi employed teachers and it was found that most of the teachers satisfied with their supportive work conditions/environments. Other parameters like Student’s behavior, parent-teacher relationship, administrative satisfaction [64], social culture, stress, demographic variables [5] are also important to evaluate the teacher's satisfaction. In recent research, [46] enhanced the concept of [40] using hypothesis (22no) testing using 5022 samples of Abu Dhabi employed teachers. This study results a strong bond between the parameters “Positive affect” and “Work condition”. “Goal progress” and “Self efficacy” are essential component where as goal support improves the goal performance if a teacher has high confidence in the work place.

Apart from the teachers’ job satisfaction; it is necessary to mine teachers’ research interest including interdisciplinary areas to create a knowledge hub and hence transforming to world class institutions.

In [63] presents a methodology for managing educational capacity utilization, simulating various academic proposals and ultimately building a Decision Support System (DSS) that gives a comprehensive framework for systematic and efficient management of the university resources.

### 4. TRENDS OF EDM RESEARCH DURING THE PERIOD 1998-2012

A survey on EDM for the period 1998 to 2012 is listed in table-3. The leverage points of this survey are the trends of DM Techniques, Tools, Dataset used and respective Educational outcomes.

**Table3. Literature survey on EDM trends during the period 1998-2012.**

<table>
<thead>
<tr>
<th>Author(s) and Year</th>
<th>Data Mining Technique(s)</th>
<th>Data Mining Tool(s)</th>
<th>Dataset</th>
<th>Educational Out Come</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sison,R.Shimura, M. 1998[59]</td>
<td>Classification</td>
<td>-</td>
<td>-</td>
<td>Students' behavior learning</td>
</tr>
<tr>
<td>Ingram,1999-2000[33]</td>
<td>Web Mining</td>
<td>-</td>
<td>Collected in web logs from WWW</td>
<td>Design and modification of websites using access patterns in weblogs</td>
</tr>
<tr>
<td>Ha et al. 2000[31]</td>
<td>Web Mining</td>
<td>-</td>
<td>-</td>
<td>To discover aggregate and individual paths of a learner in distance education system</td>
</tr>
<tr>
<td>Zaïane O.2002[74]</td>
<td>Association Rules Mining</td>
<td>-</td>
<td>-</td>
<td>On-line user behaviors</td>
</tr>
<tr>
<td>Brusilovsky,P. and Peylo,C.2003[56]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>To provide a more systematic view to the modern AIWBEES</td>
</tr>
<tr>
<td>Sheard et al. 2003 [60]</td>
<td>-</td>
<td>SPSS data analyzer, C++</td>
<td>WIRE website interaction data-2001</td>
<td>Student behavior learning</td>
</tr>
<tr>
<td>Authors</td>
<td>Method/Algorithm</td>
<td>Data Source/Techniques</td>
<td>Context</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------------------------------</td>
<td>------------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Baker et al. 2004</td>
<td>Bayesian knowledge tracing algorithm</td>
<td>-</td>
<td>Students with behavior analysis</td>
<td></td>
</tr>
<tr>
<td>Freyberger et al. 2004</td>
<td>Association Mining</td>
<td>-</td>
<td>To guide a search for a best fitting transfer model of student learning</td>
<td></td>
</tr>
<tr>
<td>Merceron and Yacef. 2005</td>
<td>Classification-DT, Clustering, Association Mining</td>
<td>Excel, Clementine, Tada-Ed, SODAS</td>
<td>Student/teachers' performance monitoring system</td>
<td></td>
</tr>
<tr>
<td>Conati et al. 2006</td>
<td>Statistical method-Hypothesis testing</td>
<td>-</td>
<td>An intelligent tutoring system which helps individual students</td>
<td></td>
</tr>
<tr>
<td>Kay et al. 2006</td>
<td>Frequent sequential pattern mining algorithm</td>
<td>Using TRAC System</td>
<td>Mining Pattern in a team work</td>
<td></td>
</tr>
<tr>
<td>Romero et al. 2007</td>
<td>Statistics, Visualization, Classification-DT, Clustering, Association Mining</td>
<td>WEKA, GIS MO, KEA</td>
<td>EDM-its research practice</td>
<td></td>
</tr>
<tr>
<td>Vandamme et al. 2007</td>
<td>Decision Tree, Neural Network</td>
<td>-</td>
<td>Academic performance prediction</td>
<td></td>
</tr>
<tr>
<td>Mansmann and Scholl 2007</td>
<td>Statistics, Clustering, Classification-DT, Clustering, Association Mining, Rule Mining, Visualization, web mining</td>
<td>WEKA</td>
<td>Decision Support System</td>
<td></td>
</tr>
<tr>
<td>Romero et al. 2008</td>
<td>Statistics, Clustering, Classification, Association Rule Mining, Visualization, web mining</td>
<td>WEKA</td>
<td>Mining e-learning Data</td>
<td></td>
</tr>
<tr>
<td>Delavari, 2008</td>
<td>Classification-DT, Clustering, Neural Network (mixture)</td>
<td>-</td>
<td>Decision making process</td>
<td></td>
</tr>
<tr>
<td>Jiang et al. 2008</td>
<td>-</td>
<td>-</td>
<td>EDM-its research practice</td>
<td></td>
</tr>
<tr>
<td>Perra et al. 2009</td>
<td>Clustering, Sequential pattern mining</td>
<td>TRAC</td>
<td>Importance of leadership and group interaction in educational process</td>
<td></td>
</tr>
<tr>
<td>Zurada et al. 2009</td>
<td>Classification</td>
<td>-</td>
<td>Web learning</td>
<td></td>
</tr>
<tr>
<td>Baker and Yacef. 2009</td>
<td>-</td>
<td>-</td>
<td>EDM-its research practice</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Methodology</td>
<td>Tools/Software</td>
<td>Data Source/Administration</td>
<td>Application</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>------------------------------------------</td>
<td>--------------------------------------------</td>
<td>------------------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Romero and Ventura. 2010[17]</td>
<td></td>
<td></td>
<td></td>
<td>EDM-its research practice</td>
</tr>
<tr>
<td>Al-shargabi and Nusari,2010[4]</td>
<td>Classification-DT, Clustering, Association Mining</td>
<td>WEKA</td>
<td>By survey at UST from 1993 to 2005</td>
<td>Students academic achievements, drop out and financial behavior</td>
</tr>
<tr>
<td>Jafar,2010 [49]</td>
<td>Classification, Clustering, Market Basket analysis</td>
<td>MS Excel add-in ,MS SQL Server</td>
<td>Iris and Mushrooms,UCI Repositories</td>
<td>DSS</td>
</tr>
<tr>
<td>Zhang et al. 2010 [75]</td>
<td>Naïve Bayes, SVM, Decision Tree</td>
<td>Oracle Data Miner</td>
<td>By survey 5,458 data Thames Valley University, UK</td>
<td>To improve student retention in higher education</td>
</tr>
<tr>
<td>Gupta et al.,2011 [20]</td>
<td>Classification-Decision Tree</td>
<td>SPSS Clementine</td>
<td>AIEEE 2007</td>
<td>Predicting Enrolment Decision</td>
</tr>
<tr>
<td>Dutta Borah et al., 2011 [45]</td>
<td>Classification-Decision Tree</td>
<td>SPSS Clementine 11.1</td>
<td>AIEEE2007</td>
<td>Predict student’s enrollment decision</td>
</tr>
<tr>
<td>Wang and Liao, 2011 [68]</td>
<td>ANN</td>
<td></td>
<td>By survey, University of Central Taiwan</td>
<td>Explores learning performance in an e - learning system</td>
</tr>
<tr>
<td>Alberg et al.,2012 [22]</td>
<td>Regression Tree</td>
<td></td>
<td>Online data</td>
<td>KDD support system</td>
</tr>
<tr>
<td>Lee and Chen.,2012 [29]</td>
<td>Association pattern Mining</td>
<td></td>
<td>Online data</td>
<td>Security analysis</td>
</tr>
<tr>
<td>Calders and Pechenizkiy, 2012 [66]</td>
<td></td>
<td></td>
<td></td>
<td>EDM-its research practice</td>
</tr>
<tr>
<td>Huebner,2012[57]</td>
<td></td>
<td></td>
<td></td>
<td>EDM-its research practice</td>
</tr>
<tr>
<td>Lin S.H., 2012 [61]</td>
<td>Decision Tree Algorithm</td>
<td>WEKA</td>
<td>5943 records of 1st year students of Biola University</td>
<td>Student Retention Management</td>
</tr>
</tbody>
</table>

The survey conducted by [7, 15], reported that during 1995-2005, 28% research are involved in prediction methods; 43% are relationship mining, 17% are exploratory data analysis and 15% are cluster analysis. During 2008-2009, it has been reported that only 9% researches are involved with relationship mining where as the rest are approximately same.
This survey (see table -3) considered the four groups of years i.e. 1998-2000, 2001-2004, 2005-2008 and 2009-2012 to find out the paradigm shift in EDM (see Figure 2).

Trends of DM Techniques

During 1998-2000, highest research involved with Web Mining whereas during 2001-2004 researches were involves Association Mining. This research revealed the almost same result given by [7, 15].

Changes in DM methods/techniques are seen during the periods 2005-2008 and 2009-2012. During this period, highest research involved in Classification- DT algorithms and Clustering. Association Mining begs the next position. Apart from these techniques, researches involving other DM Techniques SVM such as Neural Network etc. are seen during 2009-2012.

Trends of DM Tools Used

DM tools are required to validate the large set of data collected from heterogeneous environments [58]. During 1998-2012 it is found that researcher mostly preferred open source tool like WEKA and then commercial tool such as SPSS Clementine to validate their dataset(see figure 3).
Trends of Dataset Use

EDM researchers’ preferred Web data during 1998-2000. Whereas during the period 2001-2004 the researches preferred data from educational institutions in their research works. The uses of primary data (survey data) and secondary data (public repository data) in the research were seen during 2005-2008 and 2009-2012. It is beneficial to the EDM research community to use both the data set. But the caution should be taken to apply secondary data set as it may not be inadequate in the context of the EDM research problem [19].

Trends of Educational Outcome

Behavioral identification of learners using the web was the main focus of research during 1998-2000.

The research paper “Discovering web access patterns and trends by applying OLAP and data mining technology on web logs” by [72]mostly cites papers (citation no: 553) as on 30th January 2013 (see figure 4).

During 2001-2004, the main focus of the researches were to design an intelligent web based educational system, recommender system and educational planning in general.

The research paper “Web usage mining for a better web-based learning environment” by [73], mostly cites papers (citation no: 188) as on 30th January 2013.

The major outcomes of research during 2005-2008 was an intelligent tutoring system which identifies Meta cognitive skills of students, DSS which evaluates overall academic performance and survey on EDM.

It is worth mentioning that in Survey category papers in EDM, the research paper “Educational Data Mining: A survey from 1995 to 2005” by [15] is found to be mostly cited papers (citation no: 327) as on 30th January 2013.

The paradigm shift of the EDM research has been seen during 2009-2012. During this period the focal point of research was importance of leadership, Users’ preference mining, student-Teacher modeling, higher education planning, EDM research & practice and Security analysis.

The research paper “Educational Data Mining : A review of the state of the Art” by [[Romero and Ventura 2010], are found to be mostly cited papers (citation no: 79) as on 30th January 2013.

Figure 4. Most cited EDM papers as on 30th January 2013.
5. DISCUSSION

A This survey focused on research trends on EDM since the year 1998 to 2012 and found that maximum research focuses were on academic objectives. The other issues are:

(1) Challenges of EDM

—Educational data is incremental in nature

Due to the exponential growth of data, the maintaining the data warehouse is difficult. To monitor the operational data sources, infer the student interest, intentions and its impact in a particular institution is the main issue.

Another issue is the alignment and translation of the incremental educational data. It should focus on appropriating time, context and its sequence.

Optimal utilization of computing and human resources [28] is another issue of incremental educational data.

—Lack of Data Interoperability

Scalable Data management has become critical considering wide range of storage locations, data platform heterogeneity and a plethora of social networking sites [27].

E.g.: Metadata Schema Registry is a tool to enhance to enhance Meta data interoperability.

So there is a need to design a model to classify/ cluster the data or find relationships. Examples of clustering applications are grouped students based on their learning and interaction patterns used in [3] and grouping users for purposes of recommending actions and resources to similar users. It is possible to introduce Neuro-Fuzzy mining technique to remove the gap of data interoperability.

—Possibility of Uncertainty

Due to the presence of uncertain errors, no model can predict hundred percent accurate results in terms of student modelling or overall academic planning.

—Research Expertise Relation between Student-Teacher

In most of the higher Educational institutions (e.g. Engineering Institutions) final year students have a compulsory project work which is a research work based on their area of interest. Generally Supervisors are assigned as per availability and area of expertise in the respective department. But still it is not possible to assign all the students –supervisor with similar area of interest hence the result of the project is nots applicable to real scenarios. There is need to find the relation between areas of interest, students’ interest, applicability of the project/research and mining cross faculty interest. It will be beneficial to introduce using Association Mining to optimize this issue.

(2) Limitations of this research

This survey work studied around 50 EDM research papers from various journals/conferences of repute in the context of DM techniques/methods, Tools, citation nos, Dataset used, educational outcomes, useful commercial / open sources/ open access tools with their features, data set and links. Since it is not possible to cover all the research papers, from all corners and explores each
and every mentioned tools with their functional points, popular tools, techniques and most cited research papers were discussed which may be considered as representatives of this research area. The features discussed in this work are comprehensive rather than inclusive.

6. CONCLUSION AND FUTURE WORK

In Information Technology (IT) driven society, mining of heterogeneous data is an important issue. In this paper, a journey of research and practice from the year 1998 to 2012 is presented. This work focuses on research trends in Offline, Online and Uncertain data, useful data sources, links etc in an educational context. Different colleges/institutions affiliated to the same University should adopt a single model for academic planning to strengthen the utilization of existing resources. Lastly this work can further be improved for designing Knowledge Discovery based Decision Support System (KDDS) which will capable of giving right decision for research in Science & Technology based on the demand of the society.

As an extension of this work we will try to solve the issues of:

—Building a real model taking into consideration of specific application of Tools and Techniques

—Build a predictive model using incremental data.

REFERENCES


[38] Zurada, J. M. et al. (2009), “Building Virtual Community in Computational Intelligence and Machine Learning”, *IEEE Computational Intelligence Magazine*, pp.43-54. DOI: 10.1109 / MCI. 2008. 9309 86.


Authors:

1. Prof. Rajni Jindal

Prof. Rajni Jindal is functioning as Head, Department of IT, Indira Gandhi Technical University for Women (IGDTUW). She joined IGDTUW as Professor in June 2012. Before joining IGDTUW, she was Associate Professor in the Department of Computer Engineering of Delhi College of Engineering, Delhi and a faculty member in the same since 1992. Prior to joining teaching she has also worked with Tata Consultancy Services (TCS). She has authored 5 books including books on “Data Structures using C” and “Compiler-Construction and Design”. She has also authored/co-authored around 45 research papers in national/international journals/Conferences. She is conducting research in the field of Data Mining and Data Bases. She is the senior member of IEEE(USA), member of WIE(USA), associate life member of CSI(India) and the life member of ISTE.

2. Ms. Malaya Dutta Borah

Malaya Dutta Borah is a full time research scholar in the Department of Computer Engineering at Delhi Technological University (DTU), Delhi, India. Prior to this she worked at Delhi Technological University (formerly Delhi College of Engineering)(2007- 2012) as Assistant Professor (Contractual Appointment), Assam Engineering College, Guwahati University (2005-2006) as Guest lecturer, Central IT College, Guwahati, Assam (2005 – 2006) as lecturer and Community Information Centre, Dhakuakhana, a CIC project under the Ministry of Information & Communication Technology, Govt of India (2002-2005) as CIC Operator.

She did her M.E. (Master of Engineering- Distinction holder) in Computer Technology & Applications from Delhi College of Engineering, Delhi University. She has also authored/co-authored around 10 research papers in national/international journals/Conferences. She is actively involved in research in the field of Data Mining, Cloud Computing, ICT and e-governance. She is the associate member of CSI(India).