Educational Survey in Research Trends using Data Mining

Neelam Mary Vijaya Nirmala
Asst Professor
Department of CSE
VNR Engineering College,Ponnuru,Guntur(Dt),AP,India
Nirmala.neelam@gmail.com

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

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).

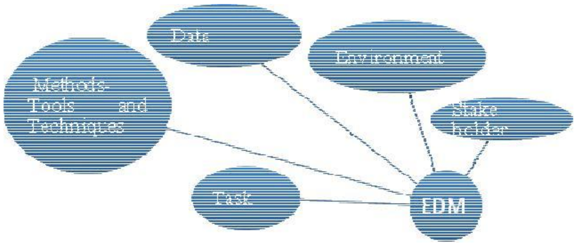

Figure1. EDM components

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, CIRCSlM-Tutor, Why2-Atlas, SmartTutor, AutoTutor, ActiveMath, Eon, GTE, REDEEM related to tutoring system.

—Collaborative learning used in [55]

—Adaptive Educational System [1]

—Learning Management System, Cognitive Learning, Recommender System used in [35] and User Modeling [18] etc.

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, Tran scripted 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.

3. MINING EDUCATIONAL OBJECTIVIES

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

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/Methods used

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.

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

[1] Amershi, S., and Conati, C., (2009) "Combining unsupervised and supervised classification to build user models for exploratory learning environments" *Journal of Educational Data Mining.*Vol.1, No.1, pp. 18-71.

[2] Mercheron, A., and Yacef, K. (2005), "Educational Data Mining: a case study" *in Proc. Conf. on Artificial Intelligence in Education Supporting Learning through Intelligent and Socially Informed Technology*. IOS Press, Amsterdam, The Netherlands, pp. 467-474.

[3] Amershi, S., Conati, C. and Maclaren, H., (2006) "Using Feature Selection and Unsupervised Clustering to Identify Affective Expressions in Educational Games", *in Proc .Of The Intelligent Tutoring Systems Workshop on Motivational and Affective Issues*. pp. 21-28.

[4] Al-shargabi, A.A. and Nusari, A. N (2010), "Discovering Vital Patterns From UST Students Data by Applying Data Mining Techniques", *in Proc. Int. Conf. On Computer and Automation Engineering,* China: IEEE, 2010, 2,547-551.DOI:10.1109/ICCAE.2010.5451653.

[5] Badri, M., and El Mourad, T (2011) "Measuring job satisfaction among teachers in Abu Dhabi: design and testing differences*" in Proc.Int. Conf. on NIE, 4th Redesigning* Pedagogy.Singapore.

[6] Baker, R.S, Corbett, A.T., Koedinger, K.R (2004) , "Detecting Student Misuse of Intelligent Tutoring Systems" *in Proc. Lecture Notes in Computer Science*Vol.3220,531-540.

[7] Baker, R.S.J.D.,and Yacef, K.(2009), "The state of Educational Data Mining in 2009:A review and future vision" *Journal of Educational Data Mining*, Vol.1,No. 1,pp.3-17.

[8] Nelson,B., Nugent, R., Rupp, A. A.(2012), "On Instructional Utility, Statistical Methodology, and the Added Value of ECD: Lessons Learned from the Special Issue", *Journal of Educational Data Mining*.Vol.4, No.1,pp.224-230.

[9] Baradwaj,B.K., and Pal,S.(2011), "Mining Student Data to Analyze Students' Performance" *.International Journal of advanced Computer Science and applications*.2,6.

[10] Aggrawal,C.C, and Yu, P.S.(2009), "A Survey of Uncertain Data Algorithm and Applications" *IEEE Transactions on Knowledge and Data Engineering*, Vol.21,No. 5,pp.609-623.

[11] Lee, C.H., Lee, G., Leu, Y.(2009), "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning" *.Expert Syst. Appl. J.Vol.*36, pp.1675-1684.

[12] Chrysostomu K. el al.(2009), "Investigation of users' preference in interactive multimedia learning systems: a data mining approach",*Taylor and Francis online journal Interactive learning environments*. Vol. 17,No. 2.

[13] Conati,C., Muldner,K., and Carenini G.,(2006)., ".From Example Studying to Problem Solving via Tailored Computer-Based Meta-Cognitive Scaffolding: Hypotheses and Design", *Technology, Instruction, Cognition, and Learning - Special Issue on Problem Solving Support in Intelligent Tutoring System.*Vol.4, No.1-54.

[14] Cocea,M., and Weibelzahl,S (2009), "Log file analysis for disengagement detection in e-learning environments", *Springer Journal User Modeling and User Adapted Interaction.* Vol.19, No.4, pp.341-385. DOI: 10.1007/s 11257-009-9065-5.

[15] Romero, C., and Ventura, S. (2007), " Educational Data Mining : A survey from 1995 to 005" *Expert Systems with Applications.* Vol. 33, pp.135-146.

[16] Romero,C. et al.,(2008), "Data Mining in course management systems: Moodle case study and tutorial", *Computer and Education, Elsevier publication. Vol.* 51, No. 1,pp.368-384.

[17] Romero, C., and Ventura, S. (2010), " Educational Data Mining: A review of the state of the Art", *IEEE Trans.on on Sys. Man and Cyber.-Part C: Appl. and rev.*, Vol.40, No.6, pp. 601-618.

[18] Romero, C., and Ventura S.(2013), " Data Mining in Education". *WIREs Data Mining and Know.Dis.,* Vol.3,pp.12-27.

[19] Kothari, C.R,(2004)*Research Methodology Methods and Techniques*, 2nd International Publishers, New Delhi,pp.99-111.

[20] Gupta,D., Jindal,R., Dutta Borah,M (2011), "A Knowledge Discovery based Decision Technique in Engineering Education Planning" *in Proc. Int. Conf. On Emerging Trends & Technologies in Data management.* Institute of Management Technology Ghaziabad, pp. 94-102.

[21] Dalip, D.H., Gonclaves, M. A. (2011), "Automatic Assessment of Document Quality in Web collaborative Digital Libraries",*ACM Journal of Data and Information Quality.*Vol.2,No.3, pp.14.DOI 10.1145/2063504.2063507

[22] Alberg, D., Last, M., and Kandel, A (2012), "Knowledge discovery in data streams with regression tree methods" *John Wiley & Sons, Inc,*2,pp.69-78,DOI:10.1002/widm.51

[23] Dey,D., and Sarkar,S.(2002) "Generalized Normal Forms for Probabilistic Relational Data" *IEEE Transactions on Knowledge and Data Engineering*, Vol.14,No.3,pp. 485-497. DoI:10.1109/TKDE.2002. 1000338.

[24] Perera,D. et al.(2009), "Clustering and sequential pattern mining of online collaborative learning data", *IEEE Transactions on Knowledge and Data Engineering,Vol.*21, No.6,pp.759-772.

[25] Shangping,D. and Ping,Z.(2008), "A data mining algorithm in distance learning.",in *Proc. Int. Conf. Comput. Supported Cooperative Work in Design.* Xian, China,pp.1014-1017.

[26] Freyberger,J., Heffernan, N., Ruiz, C.(2004), "Using

association rules to guide a search for best fitting transfer models of student learning", *Workshop on Analyzing Student-Tutor Interactions Logs to Improve Educational Outcomes at ITS Conference.*

[27] Deka,G.C(2013), ".A survey on Cloud Data Base", *IEEE Transaction on IT professional*, pp.99,DOI : 10.1109/MITP.2013.1

[28] Deka, G.C. ,and Dutta Borah, M. (2012), "Cost Benefit Analysis of Cloud Computing in Education", *in Proc. Int. Conf. On Computing, Communication and Applications.pp.*1-6.

[29] Lee, G.,and Chen, Y.C.(2012), "Protecting sensitive knowledge in association pattern mining", *John Wiley & Sons, Inc* .2, pp.60-68.,DOI:10.1002/widm.50.

[30] Dekker, G., Pechenizkiy, M., and Vleeshouwers J.(2009), "Predicting students drop out: A case study", *In Proceedings of the 2nd International Conference on Educational Data Mining, pp.*41-50.

[31] Ha, S., Bae, S., and Park, S. (2000) "Web mining for distance education" in *Proc. Int. Conf. On Management of Innovation and Technology,* IEEE. Pp.715-719.

[32] Chu,H. C.,Hwang, G. J.,Wu, P. H., and Chen, J. M. A (2007), "Computer-assisted collaborative approach for E-training course design",in *Proc. Conf. Adv. Learn. Technol*, Niigata, Japan: IEEE, pp.36–40.

[33] Ingram, (1999-2000), "Using Web Sever Logs in Evaluating Instructional web sites", *Journal of Educational Technology Systems*. Vol.28,No.2.

[34] Nesbit, J. C., Xu, Y., Winne, P. H., Zhou, M., (2008), "Sequential pattern analysis software for educational event data", in *Proc. Int. Conf. Methods Tech. Behav. Res*. Netherlands,pp.1-5.