Introduction to The Special Section on Educational Data Mining

Toon Calders
Department of Computer Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
t.calders@tue.nl

Mykola Pechenizkiy
Department of Computer Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
m.pechenizkiy@tue.nl

ABSTRACT

Educational Data Mining (EDM) is an emerging multidisciplinary research area, in which methods and techniques for exploring data originating from various educational information systems have been developed. EDM is both a learning science, as well as a rich application area for data mining, due to the growing availability of educational data. EDM contributes to the study of how students learn, and the settings in which they learn. It enables data-driven decision making for improving the current educational practice and learning material. We present a brief overview of EDM and introduce four selected EDM papers representing a crosscut of different application areas for data mining in education.

1. INTRODUCTION

Recently, the increase in dissemination of interactive learning environments, learning management systems (LMS), intelligent tutoring systems (ITS), and educational hypermedia systems as well as the wider use of ICT in education in general has allowed the collection of huge amounts of data. The increase in instrumented educational software, as well as state databases of student test scores, created large repositories of data reflecting how students learn. Some examples of popular systems include: general purpose LMS such as Sakai[1] and Moodle[2], specialized ITSs like the Cognitive Tutors[3] or SQL Tutor[4], professional education and training systems such as simulators, systems for learning elementary skills; for instance reading and performing arithmetic operations such as Neure and Ekapeli[5], and eHealth and patient education such as Philips Motiva[6]. Educational Data Mining aims at discovering useful information from the large amounts of electronic data collected by these educational systems. EDM as an emerging multidisciplinary research area brings together researchers and practitioners from computer science, education, psychology, psychometrics, and statistics.

EDM as a separate research field started to mature a few years ago. The Educational Data Mining International Conference series was launched; the 4th edition of the conference was held this year in Eindhoven, the Netherlands [6]. In 2010 the KDD Cup at the ACM SIGKDD Conference was devoted to the Educational Data Mining Challenge[7] - the task was to predict student performance on mathematical problems based on data from logs of student interaction with an ITS. The web portal of the International Educational Data Mining Society[8] provides pointers to the main resources and scientific events in this field. In the EDM area there are not as many benchmarks as in data mining or information retrieval. Student enrollment data and LMS data is rarely anonymized and made publicly available. The most known repository for data on the interactions between students and ITS educational software is maintained by the Pittsburgh Science of Learning Center (PSLC) DataShop [4]. Next to data, the repository also includes a suite of tools to process, explore and visualize the data through a web-based interface.

Historically, the majority of the EDM researchers has a background in ITS, AI in education (AIED), user modeling, technology enhanced learning (TEL), or adaptive educational hypermedia. Relatively few scientists come with a data mining background. The goal of this special section is therefore twofold: providing an overview of the field and also attracting interest from the Data Mining community. We feel that EDM can and should attract further attention of the KDD community. With this introduction to the special section we attempt to answer the question—"What is interesting in EDM for the Data Mining and KDD community?" We discuss the landscape of EDM applications and tasks in Section 2, pointing to different kinds of data available for mining, and introduce four papers selected to represent the current state of the art in the field in Section 3. Section 4 concludes the introduction.

2. TYPICAL EDM TASKS

Figure 1 presents the basic setting of EDM having a few groups of stakeholders (learners, teachers, study advisers, directors of education, educational researchers) who can benefit from EDM in different ways. For instance, students can receive advice and recommendations about available

[1] http://sakaiproject.org
[2] http://moodle.org/
[3] http://pact.cs.cmu.edu/
[4] http://www.cosc.canterbury.ac.nz/tanja.mitrovic/sql-tutor.html
[5] http://www.lukimat.fi/
[6] http://www.healthcare.philips.com/main/products/telehealth/products/motiva.wpd
[7] https://pslcdatashop.web.cmu.edu/KDDCup/
[8] http://www.educationaldatamining.org/

courses, learning activities, resources, or tasks that are the most suitable w.r.t. their current knowledge and learning objectives; teachers can see how effective their learning material is, how well the students are doing on particular tasks, and how informative test assignments are; a study adviser can identify risk groups among the students; directors of education can see how the students actually study and what the bottlenecks are in the current curriculum. In either case it is expected that the mined knowledge can give a better insight, facilitate and enhance the educational processes and the learning as a whole. The educational data mining survey by *Romero and Ventura* [8] provides an elaborate overview of how different EDM stakeholders can benefit from mining various educational data sources, and several success stories can be found in the first Handbook on EDM [9].

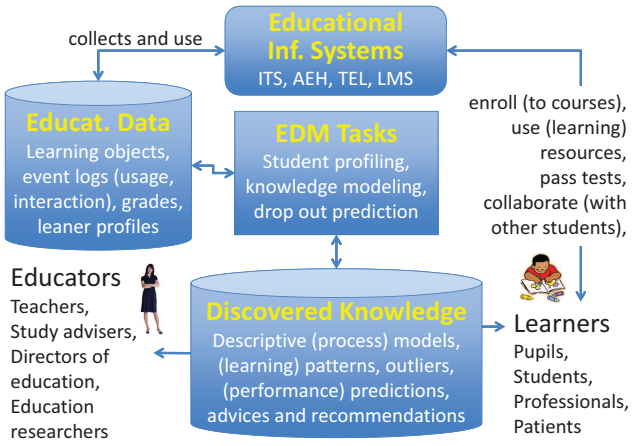

Figure 1: Educational data mining in a nutshell.

The current mainstream EDM research is primarily focused on mining ITS and LMS logs. However, EDM in a wider perspective is aimed at helping to address problems related to different phases in the leaning process, whether it is formal (e.g. tests) or informal (e.g. educational games), intentional (e.g. tutoring) or unexpected (e.g. using the social media). Examples of particular problems include:

- How to (re)organize the classes, or assessment, or placement of materials based on usage and performance data.
- How to identify those who would benefit from provided feedback, study advice or other help.
- How to decide which kind of help, feedback or advice would be most effective.
- How to help learners in finding and searching useful material, individually or in collaboration with peers.

Available Data Sources. Different kinds of information systems are supporting educational processes at different levels. For instance, administrative databases store enrollment information; i.e., who follows which program, takes which courses and (re-)exams, the student demographics and their pre-university data, such as school grades. LMSs store more fine-grained data including resource usage logs (e.g. handouts, videorecordings), assessment data, collaborations in wikis or versioning systems, and participation in forums. ITSs and educational games often have learners' performance data over a large collection of learning taks. Consequently, learning-related data may have varying characteristics. In traditional education, faculty or university level data is longitudinal (including exams data over 5-year study programmes) but corresponds only to a few hundred or thousand students. In e-learning the use of widely accepted ITSs like SQL tutor or some of the Carnegie Learning[9] tutoring tools used in schools at the national level in the United States resulted is huge datasets containing long sequences of learners' actions and their correctness. It is typical to assume that the knowledge of learners increases and skills improve over time and need to be modeled and traced. In general, the data can be seen and modeled at different levels of aggregation.

EDM Problem Formulations. A lot of basic EDM tasks can be mapped to traditional data mining problem formulations:

- *Classification*: categorizing and profiling students, determine their learning styles and preferences [1].
- *Predictive modeling*: inducing models that can predict whether (and when) a student will pass a course or not [3], will eventually graduate or drop out [2].
- *Clustering*: grouping similar students (based on behavior, performance, etc) or grouping similar courses, assignments, etc together, exploring collaborative learning patterns [7].
- *Biclustering*: finding which questions (tasks, courses, etc) are difficult/easy for which students.
- *Frequent pattern mining*: finding (elective) courses often taken together or popular paths in study programs or actions in LMS [10].
- *Emerging pattern mining*: finding patterns that capture significant differences in behavior of students who graduated vs. those students who did not or that explain the changes in behavior of student generations over different years.
- *Collaborative filtering and recommendations*: recommending suitable learning objects, based on the analysis of the performance of other learners, recommending remedial classes to students [5].
- *Visual analytics*: facilitating reasoning about the educational processes or learning results via interactive data/model visualization, e.g. visualizing collaborations of students.
- *Process mining*: understanding the study curriculum, how students follow it, (not) obeying particular constraints, understanding bottlenecks in particular study programs.

Some of the state-of-the-art data mining techniques already have been shown to be useful in particular educational domains. However, many other EDM-related areas still remain unexplored.

[9] `http://www.carnegielearning.com/`

3. CONTRIBUTED ARTICLES

To illustrate the current state of the EDM field, we have selected four contributions that together provide an overview of the main research directions in EDM. The goal of this special section on EDM is by no means to be exhaustive, yet to provide a crosscut of the field. There have been numerous other nice contributions in the field, many of which can be found in the proceedings of the past and upcoming EDM[10], ITS[11], and AIED[12] conferences and in the JEDM[13] and UMAUI[14] journals, among others.

In this special section we included the following four papers:

- **Data Mining for Improving Textbooks** by *Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi*. This paper discusses various ways for assessing the quality of existing textbooks, as well as for suggesting additional material, such as illustrations or Wikipedia pages. The quality assessment of the textbook sections is not only based upon a textual analysis of, e.g., average word and sentence lengths, but also includes an elaborated analysis of the concepts in the text and their relations. Based upon the concept graph, the dispersion of the book section is measured. In the process of analyzing the texts and suggesting additions, a lot of external information sources are used and combined, including synsets from Wordnet, and pages from Wikipedia pages with their revision history. This paper is a nice example of how a creative combination of existing techniques with the wealth of available online material allows for new applications in the educational field that were previously impossible.

- **Social Network Analysis and Mining to Support the Assessment of On-line Student Participation** by *Reihaneh Rabbany, Mansoureh Takaffoli and Osmar R. Zaïane*. Next to the study material, also the way students use it and discuss about it can be analyzed. Many electronic learning environments such as Moodle, Blackboard and others offer tools for students to collaborate. A popular example of such a collaborative tool is a forum in which students can post questions and remarks, and react to each other's contributions. Nevertheless, as *Rabbany et al.* argue, it is often quite difficult to analyze in what way students are using these tools, how they are collaborating, and what topics they are discussing about. Therefore, *Rabbany et al.* present their Meerkat-ED toolbox for social network analysis in the context of the assessment of student collaborations and course participation. The visualizations include the visualization of detected communities among the students, of keywords representing discussion topics and their relations, and the relative centrality of students in the discussions. A case study for one course is presented.

- **Mapping Question Items to Skills with Non-negative Matrix Factorization** by *Michel C. Desmarais*. Another important source of information in the educational process are the test scores of students. *Desmarais* shows how the scores of different students on a set of questions can be used to determine the skills required for a particular question, and how strong the different students are for these skills. *Desmarais* applies matrix factorization techniques for this purpose. The student-question score matrix is decomposed into two matrices: one students-skills and one skills-questions matrix. Given the constraints of the domain, non-negative matrix factorization is used; i.e., it is assumed that the skill mastery level of the students is non-negative and being more skilled will never have a negative impact on the student's ability to answer a question correctly. *Desmarais* studies the capabilities and limitations of this technique and illustrates them on two real datasets and on simulated data. The performance of the technique is measured as how good it clusters the questions according to a pre-defined categorization.

- **The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software** by *Zachary A. Pardos, Sujith M. Gowda, Ryan S.J.D. Baker, and Neil T. Heffernan*. Another example of analyzing test results is given by *Pardos et al.* In contrast to *Desmarais*, however, whose focus was mainly on detecting the required skills for different questions, *Pardos et al.* concentrate on the knowledge level of the students, and this knowledge is assumed to be non-static. The assumption is that students who solve problems evolve their knowledge, and a better knowledge will allow them to improve their performance on further questions. Knowledge about a topic, however, can be observed only indirectly through the scores of the student on questions for this topic. The knowledge of a student on a topic is therefore identified with the probability that the student will answer the next question on that topic correctly. In this way, the performance of the knowledge models can easily be assessed in controlled settings. Several models for assessing the evolving knowledge level of the students are presented, and it is shown how their predictions can be combined in ensemble methods to further boost their performance.

4. CONCLUDING REMARKS

EDM took-off. The years to come will show how this field evolves, and how it will be perceived by the KDD community – will it be yet another application domain of data mining or does it have the capacity to grow into a new subfield with its own challenges for data mining and multidisciplinary research, alike it happened for bioinformatics?

In this special section we present the current state of the art in the area inviting four representative papers, including: the evaluation and improvement of study material; assessing the knowledge of studentes based upon how they score on a set of questions; analyzing the required skills for different questions, based upon how students answer them; and visualizing collaborations of students in order to detecting groups of topics and clusters of students.

We hope you will enjoy reading the papers on EDM included in this special section and find an inspiration for formulating new data mining problems or try out your own favorite data mining algorithm on the available EDM datasets.

5. ACKNOWLEDGEMENTS

We would like to thank all the authors who contributed to this special section.

[10] http://www.educationaldatamining.org/EDM2012/
[11] http://its2012.teicrete.gr/
[12] http://www.aied2011.canterbury.ac.nz/
[13] http://www.educationaldatamining.org/JEDM/
[14] http://www.umuai.org/

6. REFERENCES

[1] H. J. Cha, Y. S. Kim, S. H. Park, T. B. Yoon, Y. M. Jung, and J.-H. Lee. Learning styles diagnosis based on user interface behaviors for the customization of learning interfaces in an intelligent tutoring system. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006*, volume 4053 of *Lecture Notes in Computer Science*, pages 513–524. Springer, 2006.

[2] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers. Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining, EDM'09*, pages 41–50, 2009.

[3] W. Hämäläinen and M. Vinni. Comparison of machine learning methods for intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006*, volume 4053 of *Lecture Notes in Computer Science*, pages 525–534. Springer, 2006.

[4] K. Koedinger, R. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press, Taylor&Francis, 2010.

[5] Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee. Targeting the right students using data mining. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'00, pages 457–464, New York, USA, 2000. ACM.

[6] M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper, editors. *Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands, July 6-8, 2011*, 2011.

[7] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):759–772, 2009.

[8] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Application*, 33:135–146, July 2007.

[9] C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker. *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press, Taylor&Francis, 2010.

[10] O. R. Zaïane. Web usage mining for a better web-based learning environment. In *Proceedings of the Conference on Advanced Technology for Education, Banff, Alberta*, pages 60–64, 2001.