Educational Data Mining: A Systematic Review of the Published Literature 2006-2013

Muna Al-Razgan[1], Atheer S. Al-Khalifa[2], Hend S. Al-Khalifa[3]

[1,3] Information Technology Department, College of Computer and Information Sciences, King Saud University
[2] Computer Research Institute, King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia
[1,3]{malrazgan, hendk}@ksu.edu.sa, [2] aalkhalifa@kacst.edu.sa

Abstract.
Educational Data Mining (EDM) is a multidisciplinary field that covers the area of analyzing educational data using data mining techniques. Since 2008 the first annual educational data mining conference has been established. Many articles have been published in the field of EDM due to the eager interest in improving teaching practices for both the learning process and the learners. This paper presents a systematic review of the published EDM literature during 2006-2013 based on the highly cited paper in this domain. More than three hundred papers were collected through Google scholar index, then they were classified according to the application domains, while also providing quantitative analysis of publications according to publication type, year, venue, category and tasks and contributors.

Keywords: educational data mining, data mining, systematic review

1 Introduction

Due to the growing availability of educational data and the need to analyze huge amounts of data generated from the educational ecosystem, Educational Data Mining (EDM) field has emerged. EDM is a multidisciplinary field that covers the area of analyzing educational data using data mining techniques (Crist´ obal Romero & Sebasti´ an Ventura, 2010). EDM is the field that "concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" as defined in the known community website [Educationaldatamining.org].

The use of e-learning and educational software in educational systems generate huge amount of data found in web servers and access logs and collected automatically as data repositories (C. Romero & Ventura, 2007). To a large degree, EDM is the analysis of student-computer interaction using tuned data mining techniques. EDM is the

T. Herawan et al. (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7_80,

modification of data mining techniques to fit the need of educational setting to help explore and discover student learning and the setting in which they learn. (Crist´ obal Romero & Sebasti´ an Ventura, 2010)

The data repository is collected from various resources such as log files, quizzes, interactive exercise, discussion forum, demographic data (gender, age, and student's grades), student's behaviors, administrative data (school, teacher, region), and many other. In addition, these data have hierarchy level such as course, subject, topics, time of access, time of observation (semester, year), level (school, college), etc. This raw educational data needs to be converted into useful information; If handled very well, it will help the educational institute improve the teaching for both teachers and students and it will have great impact on the educational research (Cristobal Romero & Ventura, 2013).

EDM seeks to use educational data repositories in understanding the learning techniques, thus building computational models that combines both the data and the theory to improve the teaching practice (Crist´ obal Romero & Sebasti´ an Ventura, 2010). In addition, EDM can find interesting information from educational data that can guide educators to established pedagogical basis when designing or modifying the course material or the learning environment to better suit the learners (C. Romero & Ventura, 2007). EDM extends the ability of traditional teacher observation to enable academics to build models based on students' attributes in real-time (R. S. J. D. Baker & Yacef, 2009). Moreover, EDM should capitalize on the known fact of Data Mining (DM) that turns data into golden knowledge, which guides academic to reach the correct decision for the learning environment (C. Romero & Ventura, 2007).

Research and contributions in the area of EDM have grown from workshops in various conferences to the establishment of EDM communities. Moreover, in 2008, the first annual international conference on EDM was held and extended to a Journal of EDM (R. S. J. D. Baker & Yacef, 2009). In the following years, EDM articles have increased rapidly, therefore, several surveys were conducted to monitor the progress of this domain, among them is a recent survey by (Cristobal Romero & Ventura, 2013) to review the state of the art in this field.

Our contribution in this paper is in twofold: (1) we propose an empirical method for surveying the literature based on the highly cited paper in Google scholar, and (2) we classify academic literature relating to EDM using a systematic review. Our paper will address the following research questions:

1. What types of applications were conducted in the field of EDM from 2007 until now?
2. What is the distribution of publications following the first survey paper by (C. Romero & Ventura, 2007)?
3. What are the classifications of published papers following (Crist´ obal Romero & Sebasti´ an Ventura, 2010) educational tasks?
4. What are the characteristics of the current research on EDM?

To answer these questions our paper is organized into two parts: the first part presents our method of data collection and selection criteria as well as data classification and analysis. The second part concludes the paper by answering our research questions and discussing the study limitations.

2 Method

2.1 Data Collection and Selection Criteria

The area of EDM has been studied extensively in many surveys published in journals and book chapters. The first published survey was published by (C. Romero & Ventura, 2007), then another theoretical paper was presented by (R. S. J. D. Baker & Yacef, 2009), afterwards an extended review paper of Romero's work was published recently (Cristobal Romero & Ventura, 2013).

In order to contribute to the area of EDM surveys, we applied a different approach for analyzing the literature. Basically, we analyzed different emerging trends in EDM using the chaining technique of (Gao, et al, 2012). Previous survey papers such as (R. BAKER & K. YACEF, 2009), followed another approach for distilling trends by referring to the most influential papers mentioned in (C. Romero & Ventura, 2007). Their selection was based on the citations number of each paper, which was categorized to different types of EDM methods. Then they compared the classification with new types' distribution showed solely in the publications of the following two years (2008 and 2009) of the EDM conference proceedings.

However in this work we pursue a different approach, we selected the most cited paper in EDM literature based on Google Scholar's search engine using the term "educational data mining", which revealed Romero and Ventura's (2007) survey paper. This survey gave us a comprehensive list of papers, published between 1995 and 2005, which are considered as educational data mining publications by a prominent pair of authors in EDM (Baker 2009). Moreover, it is the most influential paper as being cited by 410 other papers at the time of conducting this study.
A bibliometric and qualitative analysis have been done on Romero and Ventura's paper based on the following criteria:

1- Use forward chaining of publication referencing (C. Romero & Ventura, 2007) to identify all related papers.
2- Select English academic papers published in journals, conferences, workshops, book chapters, etc.
3- Select papers with research focus on data mining in educational settings, excluding any paper on education or data mining alone, and classifying them as relevant if their primary focus was using data mining techniques in the context of educational setting.

For each paper in the citation, the authors had to read the Title and Abstract and sometimes skim the paper in order to decide whether to include it or not. Papers that were written in another language rather than English or did not have EDM as their

main focus were excluded. The data collection process resulted in 281 unique papers out of 410 and was used in this systematic review.

In the following sections, we consider how EDM has recently evolved, and investigate some of the major trends in EDM research. In order to investigate what the trends are, we analyze what researchers were studying previously, and what they are studying now, towards understanding what is new and what emerged attributes in current EDM research.

2.2 Data Classification and Analysis

After we had collected the papers, we had distributed the work among the research team as follows: The first author looked for the citation of the papers and prepared the list of statistics (number of papers published yearly, active authors, and active publications venues); this will be presented in the following sections. The second author identified the common application domains across the 281 published papers.
Then the second author carried the task of classifying the papers according to identified application domains. Application domains and their definitions will be illustrated next. Due to the large number of papers, the authors split the process of classification among them based on (Crist´ obal Romero & Sebasti´ an Ventura, 2010). The eleven categories mentioned in (Crist´ obal Romero & Sebasti´ an Ventura, 2010) were followed to classify the repository of papers collected. All of the above analysis will be presented in the following sections.

Quantitative Analysis

Articles were classified quantitatively according to the publication year, venue and type of publication. The first survey paper on EDM was published in 2006, and became since then a foundation stone to other works in the field. The number of papers about EDM has grown greatly, which can be seen in (figure 1) exceeding 70 published papers in the year 2012. Most of the publications were in the form of journal articles (121 articles), followed by 118 conference proceedings (including two posters), 14 dissertations and 28 other types of publications.

Table 1 presents the five most published authors in the repository; Sebastián Ventura has the highest number of publications with 19 published works, followed by Cristóbal Romero (16 publications). Ventura and Romero have shared many publications as co-authors along with different researchers, and it can be seen that the following three authors: Valsamidis, Kontogiannis and Kazanidis had also been co-authors in many of the listed publications.

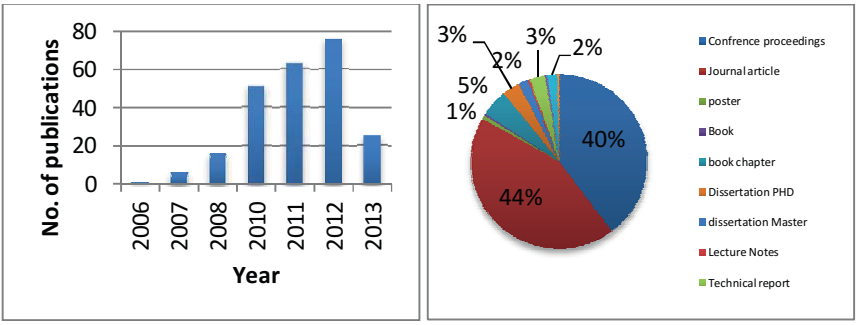

Fig. 1. Articles by publication year (left), Types of publications (right)

Table 1. Most Contributing authors

Author	No. of Articles	Details (Co-authors, year)
Sebastián Ventura	19	Porras, Romero Hervás & Zafra 2006; Romero, Antonio & De Bra 2007; Romero, Espejo & Hervás 2008; Romero & García 2008; Santos, Freire & Romero 2008; García, Romero & De Castro 2009; Romero, Zafra & De Bra 2009; García, Romero & De Castro 2009; González, Romero, del Jesús & Herrera 2009; Romero 2010; Carmona, González, Romero & del Jesús 2010; Zafra 2010; Romero & Zafra 2011; García, Romero De Castro 2011;Carmona, González,Del Jesus 2011; Zafra 2012; Espejo, Zafra, Romero & Romero 2013; Romero 2013; Márquez & Romero 2013
Cristóbal Romero	16	Porras, Ventura, Hervás & Zafra 2006; Ventura, Antonio & De Bra 2007; Ventura, Espejo & Hervás 2008; Ventura & García 2008; Santos, Freire & Ventura 2008; García, Ventura & De Castro 2009; Ventura, Zafra & De Bra 2009; García, Ventura & De Castro 2009; González, Ventura, del Jesús & Herrera 2009; Ventura 2010; Carmona, González, Ventura & del Jesús 2010; Ventura & Zafra 2011; García, Ventura, De Castro 2011; Espejo, Zafra, Romero & Ventura 2013; Ventura 2013; Márquez & Ventura 2013
Stavros Valsamidis	8	Kazanidis, Theodosiou & Kontogiannis 2009; Kazanidis, Kontogiannis & Karakos 2010; Kazanidis, Kontogiannis & Karakos 2010; Kontogiannis & Karakos 2011; Kazanidis, Kontogiannis & Karakos 2011; Kazanidis, Kontogiannis & Karakos 2011; Kazanidis, Kontogiannis & Karakos 2012; Kazanidis, Kontogiannis, Theodosiou & Karakos 2012
Sotirios Kontogiann	8	Valsamidis, Theodosiou & Kazanidis 2009; Valsamidis, Kazanidis & Karakos 2010; Valsamidis, Kazanidis & Karakos 2010; Valsamidis,

is		Valsamidis & Karakos 2011; Valsamidis, Kazanidis & Karakos 2011; Valsamidis, Kazanidis & Karakos 2011; Valsamidis, Kazanidis & Karakos 2012; Valsamidis, Kazanidis, Theodosiou & Karakos 2012
Ioannis Kazanidis	7	Valsamidis, Theodosiou & Kontogiannis 2009; Valsamidis, Kontogiannis & Karakos 2010; Valsamidis, Kontogiannis & Karakos 2010; Valsamidis, Kontogiannis & Karakos 2011; Valsamidis, Kontogiannis & Karakos 2011; Valsamidis, Kontogiannis & Karakos 2012; Valsamidis, Kontogiannis, Theodosiou & Karakos 2012

Publications' Application Domains

The articles embraced a wide-range of applications relating to EDM, the authors had identified the most used applications among the 281 articles and came up with the following list: Educational games, learning object, Mobile learning, personalized learning, scientific research into learning and learners, and intelligent tutoring system.

Classifying the articles in the previous list was done following this approach: reading the title of the paper and the listed keywords for each article then assigning the article to the appropriate application domain. Given the fact that some articles will fall in multiple application domains, we tried to match it and assign each article to one application domain.

Table 2. Types of application domains

Application Domain	Definition	# of assigned papers	% of application domains
Educational games	"Are games explicitly designed with educational purposes, or which have incidental or secondary educational value. All types of games may be used in an educational environment. "[2]	7	2%
Learning object	"a collection of content items, practice items, and assessment items that are combined based on a single learning objective"[3]	84	30%
Mobile Learning	"Learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies."[4]	5	2%

Personalized Learning	"is the tailoring of pedagogy, curriculum and learning environments to meet the needs and aspirations of individual learners"[5]	71	25%
scientific research into learning and learners	Applying educational data mining to answer questions in any of the three areas of: student models, domain models, and pedagogical support that have broader scientific benefits.(R. S. J. Baker, 2010)	95	34%
Intelligent tutoring system	"is a computer system that aims to provide immediate and customized instruction or feedback to learners"[6]	19	7%

From table 2, we notice that the majority of classified articles are found under the "scientific research into learning and learners" application domain with a total of 95 articles. Following is the "learning object" application domain with 84 articles, and "personalized learning" with 71 articles. This could be due to the fact that the main concern of educational data mining as an emergent discipline for exploring data is to explore methods to support learning and teaching processes. Nevertheless, other application domains were less common such as educational games, intelligent tutoring, and mobile learning, which suggest that more research is needed in these areas.

Publications' Tasks

In table 3, we followed the articles' categorization based on (Crist´ obal Romero & Sebasti´ an Ventura, 2010). The categories represent examples of applications or tasks in education that can be solved using data mining. Among these tasks: "Predicting student performance" is the main focus of 41 articles, followed by "Analysis and visualization of data" with 35 articles; whereas, the following categories "Detecting undesirable student behaviors" and "Developing concept maps" appeared to be the least desired tasks in the repository. However, it is important to note that many of the selected papers are included into one or more different categories. Therefore, to better highlight different emerging categories, authors were compelled to choose the most noticeable choice of category in the paper. Table 3 below lists the categories, along with their description and number of articles assigned to each category.

Table 3. Publications categories was borrowed from (Cristobal Romero & Ventura, 2013)

Tasks	Description	# of Articles
Predicting student Performance	To estimate the unknown value of a student's performance, knowledge, score or mark	41
Providing feedback for supporting instructors	To provide feedback to support educators in decision-making about how to improve students' learning and enable them to take appropriate proactive and/or remedial action	30
Recommending to students	To make recommendations to students with respect to their activities or tasks links to visits, problems or courses to be done, and so forth.	28
Student modeling	To develop and tune cognitive models of human students that represent their skills and declarative knowledge	33
Grouping students	To create groups of students according to their customized features, personal characteristics, students personal learning data, and so forth	14
Constructing courseware	To help instructors and developers to carry out the construction/development process of courseware and learning content automatically	32
Planning and scheduling	To plan future courses, student course scheduling, planning resource allocation, admission and counseling processes, developing curriculum, and so forth	18
Detecting undesirable student behaviors	To monitor students' learning progress for detecting in real time undesirable student behaviors such as low motivation, playing games, misuse, cheating, dropping out, and so forth	11
Social network analysis	Aims at studying relationships between individuals, instead of individual attributes or properties	19
Developing concept maps	A concept map is a conceptual graph that shows relationships between concepts and expresses the hierarchal structure of knowledge	13
Analysis and visualization of data	Provide basic information directly from data (reports, statistics, etc.)	35

3 Discussion and Limitations

The aim of this study is to answer the research questions related to the state of EDM research based on a systematic review of the highly cited paper in the academic literature. Using forward chaining of publication referencing of the most influential paper in the field i.e. (C. Romero & Ventura, 2007) to answer our research questions, we can report the following findings.

The types of applications conducted in the field of EDM from 2007 until now, circulate around: Educational games, learning object, Mobile learning, personalized

learning, scientific research into learning and learners, and intelligent tutoring system. Our findings suggested the EDM research community focus more on educational games, intelligent tutoring, and mobile learning.

In terms of publications distributions, the published articles have increased since 2008 and the establishment of the EDM conference; as identified previously the number of published papers has exceed 70 papers in 2012.

Also, we had classified the articles using the same categorization used by (Crist´ obal Romero & Sebasti´ an Ventura, 2010). The categories represent examples of applications or tasks in education that can be solved using data mining. The most area that has many articles was scientific research into learning and learners, learning object, and personalized learning.

As for the characteristics of current research on EDM, we can say it is similar to the finding suggested by (Crist´ obal Romero & Sebasti´ an Ventura, 2010), it would be recommended if the tools developed for EDM can be embedded automatically in any learning management system or learning process. Moreover, the tools should be easy and attractive to educators to use, and institutes have to encourage their faculty to adapt it.

Our approach of using forward chaining of publication referencing (C. Romero & Ventura, 2007) might limit the study to a selected portion of papers. Thus, it would be interesting to investigate further on using other early important works that influenced EDM field. Another limitation in our study is not considering other non-English publications; we believe that they would have been an informative addition to the study.

Finally, EDM research became an area of concrete and solid science that will evolve and will contribute to the adaption of new teaching practices.

4 References:

Baker, R. S. J. (2010). Data Mining for Education. In *International Encyclopedia of Education , 7, 112-118.*.

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.*

Crist´ obal Romero, & Sebasti´ an Ventura. (2010). Educational Data Mining : A Review of the State of the Art. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, *40*(6), 601–618.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. doi:10.1016/j.eswa.2006.04.005

Romero, Cristobal, & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. doi:10.1002/widm.1075