A Comprehensive Review of Educational Data Mining

Xu Du, Juan Yang, Jui-Long Hung and Brett E. Shelton

National Engineering Research Center for E-Learning
Central China Normal University

Department of Educational Technology
Boise State University
1910 University Dr.
Boise, ID 83725, USA

National Engineering Laboratory for Educational Big Data
Central China Normal University
Wuhan 430079, China

Corresponding author: Brett E. Shelton.
Email: brettshelton@boisestate.edu

Abstract

In reviewing nearly 700 self-described studies of educational data mining (EDM) over the past decade, only 13 reviews were attempted that synthesize research literature. Within these 13, several key issues present themselves including non-systematic selection and the confounding of results by including learning analytics research. Here, a comprehensive, systematic review aims to provide information related to EDM’s bibliometrics and research trends. Results show that EDM has gained research efforts from scholars from North America, especially the United States, as part of a rising publication trend. Of 693 articles that met inclusion requirements, 70% were coded as actual data analysis. Statistics, clustering and data visualization were the most popular analytic methods in the descriptive analysis, while regression and decision tree were the most popular methods in predictive analysis studies. Most EDM research focused on students in higher education rather than teachers or students in K-12.

1. Introduction

Various forms of online learning systems have been embraced by institutions and learners, such as Learning Management systems (LMSs), Massive Open Online Courses (MOOCs), Intelligent Tutoring systems (ITS), and game-based learning systems. In the process of using these online learning systems, huge amounts of educational data can be recorded. This information is often referred to as “educational big data”, which cannot be analyzed to further provide decision-making support without the assistance of technology. To address this technology requirement, researchers have adopted Educational Data Mining (EDM, hereafter) as a solution to automate the analysis of educational big data.

The first International conference on EDM was held in 2008. Since then, more and more researchers started to devote research efforts to this area. A widely accepted definition of EDM is considered as, “computerized methods and tools for automatically detecting and extracting meaningful patterns and information from large collections of data from educational settings” (Kumar & Sharma, 2017, p. 2). The intensity of research efforts were also reflected by the number of publications, as nearly 700 journal articles and proceedings were published in the period between 2007 and 2017. Among them, 13 review articles were identified (Romero & Ventura, 2007; Baker & Kisor, 2009; Romero & Ventura, 2010; Romero & Ventura, 2013; Peña-Ayala, 2014; Papamitsiou & Economides, 2014; Salter, Joksimovic, Kovanovic, Baker, & Gasevic, 2016; Shingari, Kumar, & Khetan, 2017; Schwendimann et al., 2017; Rodriguez-Triana et al., 2017; Dutt, Ismail, & Herawan, 2017; Muthukrishnan, Govindasamy, & Mustapha, 2017). However, several aspects of these reviews could be noted. First, many reviews combined articles relating on both EDM and Learning Analytics (LA), which is problematic due to the substantial differences between research goals of these areas. Second, two important data sources, the Journal of Educational Data Mining and Proceedings of the Educational Data Mining conference, were not included. Third, the review sample size and scopes are small (the largest one only encompassed 240 articles from 2010 to the first quarter of 2013) when compared with the total number of available articles on EDM. Therefore, this review addresses these
aspects within existing review articles that cover EDM, then, it outlines a more comprehensive review technique to reveal additional findings.

2. Literature review

Based on our search, 693 articles and proceedings were published from 2007 to 2017 (major data sources will be listed later). Within this number of included articles, 13 review articles were identified published from 2007 to 2017. The following discussion focuses on the research questions and key findings extracted from these review articles.

There are 7 articles (Romero & Ventura, 2007; Baker & Kisor, 2009; Romero & Ventura, 2010; Romero & Ventura, 2013; Alqarni, 2016; Salter, Joksimovic, Kovanic, Baker, & Gasevic, 2016; Shingari, Kumar, & Khetan, 2017) that did not provide their process for data collection. Five review articles (Romero & Ventura, 2007; Baker & Kisor, 2009; Romero & Ventura, 2010; Romero & Ventura, 2013; Alqarni, 2016) introduced and reviewed EDM methods, data sources and main applications. These 5 reported that EDM methods included not only prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models, but also outlier detections, text mining and social network analysis. Traditional and computer-based learning environments (such as: LMS, ITS, AIH) are all common data collection sources. Although the application of EDM varies, predicting student performance and providing feedback for supporting instructors are the two main EDM applications. Salter et al. (2016) reviewed 40 tools frequently used for educational data mining for purpose of providing useful information for researchers interested in conducting EDM/LA research. Shingari et al. (2017) conducted a survey on the existing performance prediction methodologies and found that previous educational records, an excellent educational track and learning emotions were significant dimensions for predicting performance. However, the articles were not systematic literature review articles.

The remaining 6 literature reviews (Peña-Ayala, 2014; Papamitsiou & Economides, 2014; Schwendimann et al., 2017; Rodriguez-Triana et al., 2017; Dutt, Ismail, & Herawan, 2017; Muthukrishnan, Govindasamy, & Mustapha, 2017) provided systematic review criteria. Two review articles (Peña-Ayala, 2014; Papamitsiou & Economides, 2014) collected related papers to review EDM methods, most prolific researchers, data sources and applications. For example, Peña-Ayala (2014) collected 240 EDM works during the period from 2010 to the first quarter of 2013. The author found that 60% of those works depicted predictive models and 40% offered descriptive models; the most typical EDM tasks were classification and clustering; and K-means, expectation maximization (EM), J48 and Naive–Bayes were the top-four most deployed algorithms. The most prolific researchers such as Ryan Baker, Cristobal Romero and Sebastian Ventura, Kenneth Koedinger, Joseph Beck, etc., were also identified. The author also concluded that most EDM works were oriented toward student modeling and assessment, including performance, behavior, learning and domain knowledge. Papamitsiou and Economides (2014) examined the research strategies, learning settings, data sources, analytic methods and research topics and categorized them by investigating 40 LA/EDM papers from journals and conferences between 2008 to 2013. They found that the majority of articles were exploratory or experimental studies. Virtual Learning Environments (VLEs) or Learning Management Systems (LMSs) were most popular learning settings. The authors found that multiple data sources, such as system logs, questionnaires, interviews, and open datasets could be gathered for analysis. Classification, clustering and regression were the most commonly used analytic methods. The most popular research topics include student behavior modeling and performance prediction, the support of students’ and teachers’ reflections, and the awareness and the improvement of feedback and assessment services.

The remaining 4 review articles (Rodriguez-Triana et al., 2017; Schwendimann et al., 2017; Muthukrishnan, Govindasamy, & Mustapha, 2017; Dutt, Ismail, & Herawan, 2017) only focused on one aspect of EDM applications or the main applications of one EDM method. For example, Rodriguez-Triana et al. (2017) reviewed LA and EDM research on monitoring, awareness and reflection in blended learning by analyzing 40 journal papers from 2010 to 2015. The authors found that these studies mainly focused on formal learning, especially in higher education. Virtual learning environments and learning management systems were the main sources for data. In addition, action-related and content-related variables were the primary indicators for monitoring, awareness and reflection. The evaluation of effectiveness most commonly used a survey instrument to investigate perceived usability, usefulness or satisfaction. Only few studies evaluated the impacts on learning with the actual performance data.

Schwendimann et al. (2017) examined educational dashboards by reviewing articles for categories of learning contexts, data sources, visualizations and analysis types in both LA and EDM by reviewing 55 journal papers between 2010 to 2015. The authors found that most of the learning dashboards were designed for students’ self-monitoring and for instructors to monitor students in formal, higher education settings. The dashboards’ data sources heavily relied on behavior logs from a single LMS platform. The visualization types, similar to traditional
dashboards, utilized bar chart, line graph, table, pie chart, and network graph visualizations. In terms of analysis type, the authors revealed that most of the studies in their review were exploratory or proof-of-concept without authentic evaluations, and therefore, it was difficult to evaluate the impacts of learning dashboard on learning.

A total of 59 articles related to students’ performance predictions and published between 2012-2016 were reviewed and classified by Muthukrishnan et al. (2017). They found that the annual publications were increasing and that the regression model was the most commonly used method for predicting performance. They revealed the three major purposes of performance prediction were to predict final course grade, predict performance on the MOOC or VLE platforms and to identify struggling students who had potential to dropout. The attributes for performance prediction mainly included the student’s personal information, grades and some basic demographics. They also reported that only 19% of the articles had considered big data and only 13 articles (22%) used a feature selection method.

Another work conducted by Dutt et al. (2017) provided a systematic literature review on Educational Data clustering by analyzing 166 related articles. They found that clustering approach had great advantages in the application of student modeling, student profiling and learning style in E-learning contexts.

In general, different conclusions made by these articles also indicated the scope is too small to obtain consistent results. Considering that there are some differences between LA and EDM, this study aims to conduct a systematic review on EDM to address these research gaps.

3. Research Method

In this research, “educational data mining” was applied as a key term for gathering articles through the Web of Science database. Two additional data sources, the Journal of Educational Data Mining and Proceedings of the Educational Data Mining conference, were also combined with Web of Science EDM articles to form the raw dataset. Because the first related EDM article was published in 2007, the actual resulting search period was from 2007 to 2017. After the searching stage, 239 journal articles (185 from Web of Science and 54 from Journal of Educational Data Mining) and 454 conference papers (693 in total) were collected. Therefore, 693 articles were analyzed for the following phase. Then the procedure followed the coding scheme proposed by Du et al. (2018) to code each article in order to summarize analysis results.

4. Results

4.1 Bibliometrics

Trends of Publication

An intuitive distribution of papers from 2007 to 2017 is presented in Figure 1. Only a single paper was published in 2007, but 15 proceedings were submitted to the EDM2008 conference. The growth trend has continued since 2015 for the number of both journal papers. The whole time span can be divided into two stages, namely, smooth growth stage from 2007-2014 and disparate growth since 2015. As the number of researchers started to focus on related issues in the field of EDM, more articles have been submitted.

![Figure 1. The publication trends from 2007 to 2017.](image)
Nationalities and authors

The distribution of the top eight countries and districts for these 693 papers is presented in Figure 2. The top three countries are the USA (405), Spain (40) and Canada (33). Scholars from North America, especially the United States, are the most active researchers in the area of EDM, followed by ones from Europe. In addition, researchers from the USA show increasing research interest dedicated to EDM.

![Total EDM Publications by Country Since 2007]

First authors’ departments and research areas

A total of 693 articles were carefully examined to identify the first authors’ departments and research areas. Approximately 82% (571/693) of EDM articles have been published in educational journals or at the EDM conference. Journals in computer science were also published larger percentages (13.85% (96/693)) of EDM articles. Therefore, these two research fields had published 96.24% of the total number of EDM articles. More than half of the 693 articles were contributed by researchers from computer science.

Most prolific journals

Figure 3 shows the top 10 journals with the highest numbers of EDM publications. First, the results indicate the top 10 are educational (60%) and computer (40%) journals. Second, the Journal of EDM, with a total of 54 articles, far exceeded the number of publications published in other journals, followed by Computer & Education, with a total of 13 articles. The journals, including Educational Technology & Society, Expert System with Applications, International Journal of Artificial Intelligence in Education, Computers in Human Behavior, International Journal of Advanced Computer Science and Applications, Expert System, Internet and Higher Education and Computer Applications in Engineering Education, have published 5 to 10 EDM articles from 2007 to 2017. One notable result is that 116 (49%) out of 239 Journal articles were published by the top 10 journals. The remaining publications were distributed in more than 90 journals. Lastly, two journals (Computers & Education and Computer Applications in Engineering Education) show an increasing trend as shown in Figure 3.
4.2 Research Trends

Research type

Four codes, including review, concept or framework only, proof of concept with small scale data analysis and actual analysis, were used to reflect the analysis level of these 693 articles. An article coded as a review article aims to represent some summarized research findings via literature review, while concept or framework only articles focus on the introduction of perspectives, concepts, or frameworks. These two categories do not involve data analysis. For example, one smart recommendation system named Micro Learning as a Service (MLaaS) was proposed for purpose of delivering personalized open educational resources to satisfy learners’ just-in-time demands (Sun et al., 2017). This system was designed based on a top-down approach, and the authors mainly focused on the construction of the knowledge base which could provide guidelines for the decision-making process of a recommendation system. Because this paper introduced how to construct this system for recommendation without data analysis, it was coded as concept or framework only.

Articles coded as proof of concept with small scale data analysis usually involved the introduction of the proposed methods or frameworks, but they validate the proposed methods or frameworks with a small scale data analysis. For example, a five-layer framework based on smart computing concept was proposed in order to facilitate automated student performance evaluation in engineering institutions (Verma, Sood, & Kalra, 2017). Finally, a dataset that included 24 students’ activity data was selected to test the scalability and utility of the proposed system. One model was proposed to predict the behaviors of new students for improving the tutoring feedback in an intelligent system (Riofrio-Luzcando, Ramirez, & Berrocal-Lobo, 2017). Then, 85 students’ log data was collected to validate the proposed model, with the results showing that this model could provide good predictions and adaptive tutoring feedback for each student type.

Articles coded as actual analysis usually reported the steps of data collection, analysis methods, results, and interpretations in detail. Implications related to analysis results were also discussed in actual analysis articles. For example, 318 students’ dataset that included reading time features, textual features and context features, were analyzed based on several supervised classification algorithms to predict if students’ minds were wandering while reading (Mills & D’Mello, 2015). Analysis results showed that the mind wandering was negatively related to posttest performance which validated the effectiveness of automated mind wandering predictor. Alsheddy and Habib (2017) collected 1980 undergraduates’ personal information and academic performance data, including course grade, semester’s GPA and cumulative GPA, to build a classifier based on a decision tree algorithm for identifying low-performing students at the end of program to support instructors’ decision-making. The authors found that the first year GPA was the most important predictor.

Coding results show that 2% (13/693) of papers were coded as review articles. Approximately 28% (197/693) of proceedings and articles were concept or framework only (63 articles, 9%) and proof-of-concept (134 articles, 19%) with small scale data analysis, and 483 articles (70%) were coded as actual analysis. The results

Fig. 3. The publication trends of top 10 journals.
indicate that the majority of EDM research studies involved actual data analysis for discovering meaningful information, consistent with the EDM target. From 483 data analysis articles identified, the following sections further discuss these 483 articles from the aspects of research method and learning environment.

**Research method**

The analysis results of 483 data analysis articles show that only five articles were coded as prescriptive analysis, which aims to discover hidden issues and propose the corresponding solutions. Two hundred and thirty-two (48%) of the articles, which employed descriptive statistics, data visualization, rules mining or other unsupervised learning techniques to conduct their studies, were coded as descriptive studies. Statistics, clustering and data visualization were the most popular analytic methods in the descriptive analysis. Two hundred and forty-seven (51%) articles, which adopted supervised algorithms, were coded as predictive studies. Regression and decision tree were the most popular methods in the predictive analysis studies.

**Learning Environment**

Except for 16 data analysis articles, which did not specify their learning environments nor belong to the examined learning environments (Higher Education, K-12 or MOOCs), the remaining 467 articles were examined to identify their sample sizes and learning environments. The results showed that (1) more than 95% of these 467 articles focused on collecting students’ data for their study, and only 19 articles gathered teachers’ data for analysis; and (2) more than 60% of these 467 articles targeted studying issues from within higher education.

5. **Conclusion**

This systematic analysis reveals the development trends of EDM by reviewing related papers from 2011 to 2017. Based on the analysis results, several conclusions can be made: (1) the publication numbers of EDM indicate a rising trend in the number of journal articles; (2) scholars from North America, especially the United States, have devoted the most research effort toward EDM research; (3) researchers from computer science have published more than 50% of the EDM articles; (4) the two most prolific journals in EDM are “Journal of Education Data Mining” and “Computers & Education”; (5) the majority of EDM research studies involve actual data analysis for discovering some meaningful information, which is consistent with the EDM target; (6) statistics, clustering and data visualization are the most popular analytic methods in the descriptive analysis, while regression and decision tree were the most popular methods in predictive analysis studies; and (7) most EDM research focused on students in higher education rather than teachers or students in K-12.

**Funding**

This research was supported by the National Science Foundation of China (No.61877027) and self-determined research funds of CCNU from the colleges’ basic research and operation of MOE(CCNU19ZN003)

**References**


