IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 10, 2020, accepted April 25, 2020, date of publication May 14, 2020, date of current version July 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994561

Big Educational Data & Analytics: Survey, Architecture and Challenges

KENNETH LI-MINN ANG[1], **(Senior Member, IEEE), FENG LU GE**[2], **AND KAH PHOOI SENG**[3,4], **(Member, IEEE)**

[1]School of Science and Engineering, University of Sunshine Coast, Petrie, QLD 4502, Australia
[2]Pacific Telecom & Navigation Ltd., Hong Kong
[3]School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia
[4]Sydney Imperial Polytechnic Institute, Sydney, NSW 2000, Australia

Corresponding author: Kenneth Li-Minn Ang (lang@usc.edu.au)

ABSTRACT The proliferation of mobile devices and the rapid development of information and communication technologies (ICT) have seen increasingly large volume and variety of data being generated at an unprecedented pace. Big data have started to demonstrate significant values in higher education. This paper gives several contributions to the state-of-the-art for Big data in higher education and learning technologies research. Currently, there is no comprehensive survey or literature review for Big educational data. Most literature reviews from a few authors have focused on one of these fields: educational mining, learning analytics with discussions on one or two aspects such as Big data technologies without educational focus, social media data in education, etc. Most of these literature reviews are short and insufficient to provide more inclusive reviews for Big educational data. In this paper, we present a comprehensive literature review of the current and emerging paradigms for Big educational data. The survey is presented in five parts: (1) The first part presents an overview and classification of Big education research to show the full landscape in this field, which also gives a concise summary of the overall scope of this paper; (2) The second part presents a discussion for the various data sources from education platforms or systems including learning management systems (LMS), massive open online courses (MOOC), learning object repository (LOR), OpenCourseWare (OCW), open educational resources (OER), social media, linked data and mobile learning contributing to Big education data; (3) The third part presents the data collection, data mining and databases in Big education data; (4) The fourth part presents the technological aspects including Big data platforms and architectures such as Hadoop, Spark, Samza and Big data tools for Big education data; and (5) The fifth part presents different approaches of data analytics for Big education data. This part provides a more inclusive discussion on data analytics which is beyond traditional forms of learning analysis in higher education. This includes predictive analytics, learning analytics including collaborative, behavior, personal learnings and assessment, followed by recommendation systems, graph analytics, visual analytics, immersive learning and analytics, etc. The final part of the paper discusses social (e.g. privacy and ethical issues) and technological challenges for Big data in education. This part also illustrates the technological challenges faced by giving an example for utilizing graph-based analytics for a cross-institution learning analytics scenario.

INDEX TERMS Big data, learning technologies, educational data, learning analytics.

I. INTRODUCTION

In a world of data deluge, vast amounts of information are generated in every area of our lives with the rapid development of new technologies such as Internet, social media, Internet of Things (IoTs), cloud, smart and mobile devices. The public, commercial and social sectors also ceaselessly produce huge amounts of data in a variety of formats from different sources. The *volume*, *variety* and *velocity* (*3Vs*) of data generated daily lead to the phenomenon of Big data with the potential to further improve the values of products and services in different industries [147], [148]. One of the sectors that *3Vs* coexist in the data is the higher education and professional education industry. Educational data are captured and generated rapidly in the higher educational ecosystem which embraces different systems and platforms

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

such as course management and learning management systems (LMS), massive open online courses (MOOC), OpenCourseWare (OCW), Open Educational Resources (OER), and social media sites such as Twitter, Facebook, YouTube and personal learning environments (PLE). The scalability to data processing and analysis enable the development of new insights and valuable information from these educational data and have further shown promise in higher education to benefit academics, students and the whole education ecosystem. Since Big data and analytics is employed to draw useful insights or *values* (the $4^{\text{th}} V$) from the educational data, we use the term Big educational data to describe this emerging field. There has been growing interest in the education community to gain insights of Big educational data to improve the learning performance of students, recommend courses, analyze learning patterns, predict dropout, improve the working effectiveness of instructors and reduce administrative workload.

Big data technologies comprise of architectures and technologies which are designed to extract valuable information from very large volumes from a wide variety of data sources. Some common platforms for Big data technologies which have been developed are Hadoop, Samza and Spark. Hadoop is commonly used for the information processing of complex Big data systems and off-line processing. Samza is mainly used to address the large volumes for high rate stream data processing, and Spark is often used for off-line rapid Big data processing. In the context of Big data in education, some specific Big data architectures or frameworks [1]–[10] have been proposed for education. The authors in [1] proposed a distributed architecture for the information processing of Big education data and predicting student performance with and without sentiment analytics. The authors in [2] proposed a five-layered architecture termed the Concept Definition for Big Data Architecture for education. The authors in [3] proposed a cloud-based architecture to analyze educational data from the Moodle system in the cloud using Apache Hadoop. The authors in [4] proposed a Big data architecture for education using Spark to identify patterns of lecture data that students have taken for the year and semester. The authors in [5] proposed a logging architecture for an E-Learning Big Data Ecosystem. The authors in [6] proposed a Big data infrastructure using the Hadoop platform. The platform is deployed within the e-learning infrastructure of a laboratory. The authors in [7] proposed an architecture based on the Apache Hadoop distributed computing architecture to process the Big data of Holland vocational interest theory.

Other works on frameworks and platforms for Big education data can be found in [8]–[10]. Further details will be discussed later in the paper. Big data analytics is changing the educational industry and gives new opportunities for both learners and instructors. In general, there are three challenges for Big educational data analysis to be addressed: (1) The huge amount of data to be processed; (2) The complex and unstructured data analytics; and (3) The difficulty to find the hidden value in the Big education data in a timely manner. The authors in [153] reported on a case study applying a Big data framework towards a LMS which was conducted at the Catholic University of Murcia. The authors commented on the challenges of managing the large volume of data generated by users in the LMS and employed statistical and association rule techniques to speed up the statistical analysis of the data. In this study the size of the Big data generated by the LMS was 70GB from data sources such as student activity, learning modality (e.g. on-campus, online, and blended), number of accesses to the LMS, tools employed by students and their associated events. In the era of Big education data, educational data mining (EDM) and data analytics are becoming essential tools to address the challenges. Data mining or also termed as knowledge discovery is known for its effectiveness in discovering hidden information embedded in the educational data. A recent literature review paper on EDM can be found in [11]. This review work presented twenty years of data mining research in e-learning environments, from an educational perspective. This paper presented a wide-scale review of 525 papers where both terms of ‘‘data mining’’ and ‘‘education’’ were analyzed and used as keywords. The review included 72 papers focused on teaching-learning evaluation. The analyzed papers showed that the researches in EDM have expanded into several different sub-areas and themes.

Other literature reviews paper on EDM can be found in [12]–[18]. Learning analytics (LA) or sometimes referred to as academic analytics, and EDM are interconnected areas in education research. A recent literature review paper on EDM and LA together for 21st century higher education can be found in [19]. There are different definitions of LA from different authors. Some authors define it in terms of the use of student-generated data for the prediction of educational outcomes for tailoring education, whereas other authors define LA as a tool to help educators examine, understand and support student study behaviors and change their learning environments. A literature review of the current landscape of the usage of LA in higher education can be found in [20]. This study was based on the analysis of 252 papers on learning analytics in higher education published between 2012 and 2018. The work by [21] proposed a literature review of the LA landscape from its evolution, status and trends. The authors discussed LA as arising from a knowledge discovery paradigm to understand the learning process. The work by [22] discussed the evidence on four propositions of LA including whether LA improves learning outcomes and student retention, completion and progression. The work by [23] focused on the current research trends of LA and its limitations and methods. Another literature review focused on the use of LA in higher educational settings can be found in [24]. Up to this point, we can see that there is no comprehensive survey or review for Big educational data. Most reviews have either focused on EDM or LA from only the education aspects. There are some short papers on Big education data but they only provide short overviews of Big data in education and challenges. Therefore, there is a need of a solid review that combine all aspects in both technologies and education for Big education data. A comprehensive literature review of Big education data which emphasizes on all aspects of Big data technologies, architectures and data analytics for education is the major contribution in this paper. The literature review in this paper has been comprehensively carried out using an extensive search of the relevant databases including IEEE Xplore, Springer, ScienceDirect, ACM conference proceedings and other sources using combination of keywords such as "Big data", "Education", "Learning analytics", "Education data mining", "Learning management system", "MOOC", "immersive learning", etc. For example, when using IEEE Xplore, a search with the keyword combination of "Big data" and "Education" returned 585 journals and 1452 conference papers. Of this, recent papers most relevant to Big educational data were surveyed.

In this paper, the data sources from education platforms or systems including LMS, MOOC, learning object repository (LOR), OCW, OER, social media, linked data and mobile learning contributing to Big education data are discussed. This is followed by the data collection, data mining and databases for education. This paper also gives discussions for the technological aspects which include Big data platforms such Hadoop, Spark and Samza and Big data tools for Big education data. The Big data architectures or frameworks specifically proposed to education are reviewed and discussed in detail. The most challenging part of this paper is to present a comprehensive literature review on data analytics from both technology and education aspects and this is beyond traditional forms of analysis in education. The works on data analytics are classified into predictive analytics, learning analytics which includes collaborative and interactive learning, behavior learning, personal learning and others. Recommendation systems or recommender for education which is an emerging topic in data analytics is also presented. Other emerging analytics such as graph analytics, visual analytics, immersive learning and analytics are also included. The final part of the paper provides some experimental insights for utilizing graph analytics for a university-based learning analytics scenario. The technological and social challenges for Big data in education and insights for future direction are also discussed. The rest of the paper is organized as follows. Section II gives background information and research classifications. Section III describes the data sources from education systems that form the Big education data. Section IV reviews the data collection, mining and databases in education systems. Section V presents the technological aspects for Big education data. Section VI gives a comprehensive literature review on data analytics. Section VII discusses future challenges for Big data in education. This section also illustrates the usefulness and technological challenges faced by giving an example for utilizing graph-based analytics for a cross-institution learning analytics scenario. The paper is concluded with some comments and remarks in Section VIII.

TABLE 1. Overall classification of big educational data research.

Classification	References
Data Sources, Collection and Mining for Big Education Data	
Educational data sources:	
Learning management systems (LMS)	[25],[26],[27],[28],[29]
Massive open online courses (MOOC)	[30],[31],[32],[33]
Open educational resources (OER), OCW	[34],[35],[36]
Social media	[37],[38]
Linked data	[39],[40],[41],[42]
Educational data collection	[43],[37]
Educational databases/datasets	[44],[45],[46],[47],[48],[49],[50],[51]
Education data mining (EDM)	[11],[12],[13],[14],[15],[16],[17],[18],[19],[52],[53]
Technological Aspects for Big Education Data	
Big data platforms	[6],[54],[55],[56]
Frameworks and architectures for Big education data	[1],[2],[3],[4],[5],[6],[7],[8],[9],[10]
Data Analytics for Big Education Data	
Predictive analytics:	
Student performance prediction	[1],[59],[60],[61],[62],[63],[64],[65],[66],[67],[68],[69],[70]
Dropout prediction and academic early warning systems	[71],[72],[73],[74],[75],[76],[77],[78],[79],[80],[81]
Courses selection	[82],[83],[84]
Learning analytics:	
Collaborative and interactive learning	[85],[86],[87],[88],[89],[90],[91]
Behavior learning	[92],[93],[94],[95],[96],[97],[98]
Personalized learning	[99],[100]
Social learning	[101],[102],[103],[104]
Learning and assessment analytics using Experience API (xAPI)	[105],[106],[107],[43],[108],[109],[110]
Recommendation systems	[111],[112],[113],[114],[115],[116],[117],[118],[119],[120],[121],[122],[123],[124],[125],[126]
Graph analytics	[127],[128],[129]
Visual analytics	[130],[131],[132],[133],[134],[135],[136],[137],[138]
Immersive learning and analytics	[139],[140],[141],[142]
Social media analytics	[37]
Future Challenges for Big Education Data	
Social challenges	[144],[145],[146]
Technological challenges	[38],[143]

Note: some papers are classified into more than one category in the table.

II. OVERVIEW AND RESEARCH CLASSIFICATION

The paper first presents the overview and classification of Big educational data and analytics research as shown in Table 1 to give a concise summary of the overall scope of this paper. The research works are classified into the various categories based on the following: (1) Big educational data; (2) Technological aspects for Big data for education; (3) Data analytics for Big education data; and (4) Future challenges for Big education data. Table 1 also allows the reader to see the full landscape of the research field of Big education data.

III. DATA SOURCES FROM EDUCATION SYSTEMS CONTRIBUTING TO BIG EDUCATION DATA

Data from education systems can be found in various sources such as student information systems, student administrative

TABLE 2. Summary of survey contributions for EDM research.

Ref.	Year	Survey objectives	Remarks and comments
[11]	2018	20 years of data mining research from educational perspective.	Authors identified and classified challenges for research to improve student learner performances in e-learning environments.
[19]	2019	EDM and learning analytics in higher education.	Authors focused on four aspects: (1) computer-supported learning analytics (CSLA); (2) computer-supported predictive analytics (CSPA); (3) computer-supported behavioral analytics (CSBA); and (4) computer-supported visualization analytics (CSVA).
[15]	2015	History and application of DM techniques in educational field (traditional educational system, web-based educational system, intelligent tutoring system, e-learning).	Authors discussed concepts for EDM such as prediction, clustering relationship mining, outlier detection, text mining, social network analysis).
[12]	2007	Highlighted main DM techniques applied to e-learning environments.	Authors proposed three useful orientations for EDM research: (1) EDM research oriented towards students; (2) EDM research oriented towards educators; and (3) EDM research oriented towards academics and administrators.
[13]	2017	Systematic review on EDM focusing on clustering algorithms and its applicability in the context of EDM.	Authors proposed the term Educational Data Clustering (EDC) and reviewed different approaches for EDC (166 studies) including for e-learning, examination failure, intelligent tutor system, learning style, student modeling, student motivation, student profiling, etc.
[16]	2013	Discussion of DM techniques in order of relevance, tendencies and limitations faced by learning environments.	
[17]	2013	Survey highlighted trends and challenges of EDM from perspectives of educational actors.	Authors introduced a new perspective on the individualization and interaction between educational actors.
[18]	2014	Survey discussed researches upon behavior detection, personalization, student performance evaluation obtained by DM techniques (clustering, classification, regression).	
[52]	2013	Survey of EDM focused on student retention and evasion, recommendation systems and course administration.	Authors focused on detecting student circumvention risks through predictive models, provide custom recommendation to students by identification of needs and learning disabilities.
[53]	2014	Covered ubiquitous and pervasive data mining applied to education for fraud detection, identification of students that require special attention.	

systems, learning management systems and from library information systems. New education developments and applications of information technology together with Internet technology have led to the online education industry. Higher education institutions are increasingly offering and delivering online learning resulting in a large volume and availability of educational digital libraries, storage repositories and tools. Furthermore, enrolled students and offered courses from massive open online courses (MOOC) are becoming large and diverse, resulting in a growing abundance in data for analytics. There is also increasingly different varieties and formats of audio, video, text, and images besides the data in relational databases from institutions. This section presents sources that contribute to Big educational data by reviewing the current education systems or platforms. Fig. 1 shows a pictorial overview of research areas and data sources in Big education data.

A. LEARNING MANAGEMENT SYSTEMS (LMS)

Learning management systems (LMS) are educational management platforms for the administration, delivery, tracking

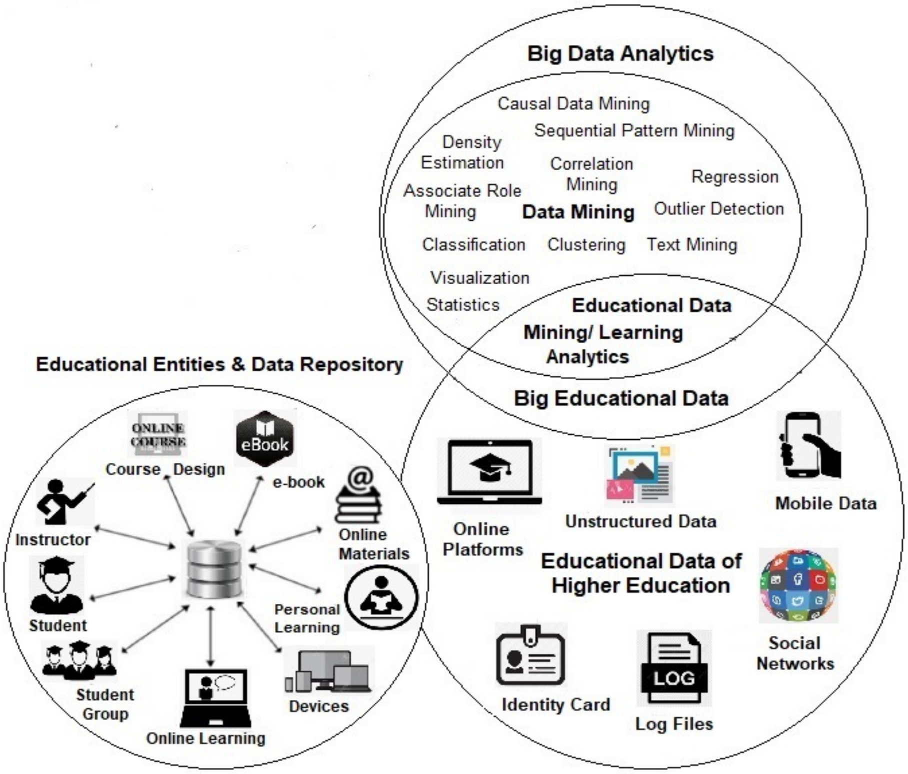

FIGURE 1. Overview of research areas and data sources for big education data.

and reporting of educational curriculum and courses. Moodle [28] is one of the most popular open source LMS options available today. Other examples of LMS [29] are Canvas [151], Sakai [152], ATutor, Eliademy. Forma LMS, Dokeos and OpenOLAT. The LMS concept emerged from e-Learning. In general, LMS have three major functions: (1) Management of educational courses and students; (2) Management of online assessments and tracking student progress and attendance; and (3) Providing feedback to users and students. The LMS provides services and tools to instructors to create course content which contains text, images, tables, interactive tests, and slideshows. The LMS can also be used to engage the student with contact tools and control access to the educational content. For instructors, the LMS enables the management of courses and modules, enrollment of students, and generation of reports on students. Most modern LMS are web-based information technology systems. With the advancement of technology, various tools and strategies can be employed for embedding content into LMS such as SCORM (Sharable Content Object Reference Model) [26], and LTI (Learning Tools Interoperability) [27].

B. MASSIVE OPEN ONLINE COURSES (MOOC)

Massive Open Online Courses (MOOC) employ web-based learning technologies to enroll large number of students worldwide. MOOC learning materials and contents can be delivered as text-based or video-based materials. Two different pedagogical approaches called c-MOOC and x-MOOC to distinguish MOOC are often used [30]. The c-MOOC emphasize the openness and networking among learners and facilitators where anyone can contribute to the contents, whereas x-MOOC are more facilitator-centric; the contents are prepared by the facilitators. Coursera [31] and edX [32] are two established MOOC. Other examples of MOOC [33] include Udacity, Duolingo, Treehouse and Google Primer.

C. OPEN EDUCATIONAL RESOURCES (OER) & OpenCourseWare

Open educational resources (OER) are educational materials that are freely available in the public domain. The OER include licensed text, media, and other digital assets that are useful for teaching, learning, and assessment. The term OER was introduced at the 2002 UNESCO Forum on Open Courseware [34]. Some examples of OER include: (1) university curriculum and courses, video lectures and assignments; (2) Interactive simulations about a specific topic (e.g. mathematics, chemistry, etc.); (3) Digital textbooks that are supported with additional learning materials; (4) Lesson plans, worksheets and learning activities; and (5) Translations and adaptations of previously-published OER. Some well-known examples of OER [35] include Khan Academy, OpenStax CNX, Open Textbook Library, Curriki, and Wikimedia Commons. OpenCourseWare (OCW) [36] is a subset of OER. OCW refers to the free and open digital publication of high-quality college and university level educational materials. Examples of OCW include MIT OCW, Johns Hopkins OCW and CORE (China Open Resources for Education).

D. SOCIAL MEDIA

Social media sites such as Twitter, Facebook and YouTube provide a platform for learners to share their educational experiences, emotions, concerns about the learning process and seek social support from peers. These digital data provide knowledge and perspectives for instructors to understand the student's experiences outside the classroom environment. The data from social-based environments can provide valuable knowledge to inform on student learning and assist institutional decision-making on interventions for at-risk students, improve education quality and increase student retention, and success [37]. The abundance and diversity of the social media data raises challenges for algorithms to capture the embedded information within the data.

E. LINKED DATA

Linked Data (LD) uses Internet technologies to create connections among data which may be stored in databases distributed across several geographic locations. LD extends the Web of Documents to a Web of Data, where data may be directly connected. LD principles and technologies are being investigated in various areas. Several studies target to use LD to solve problems of interoperability of educational data and resources. The authors in [39] presented a systematic mapping of proposals which have been adopting Linked Data to support education objectives. The authors discussed the challenges and provided a research landscape of the area. Some notable projects in the LD area are the LinkedUp project, Linked Education Cloud, and mEducator. LD technologies have the potential to drive the development of applications in the LA and EDM areas. The work in [40] describes the Learning Analytics and Knowledge (LAK) dataset which contains a five-year collection of bibliographic resources about learning analytics and educational data mining. Other examples of works for applying LD in LA can be found in [41] and [42]. The authors in [41] developed a metric to identify the relative ranking of universities worldwide based on educational Linked Data. The authors in [42] proposed using education and economic LD for analysis of school performance in Brazilian schools.

IV. DATA COLLECTION, MINING AND DATABASES IN EDUCATION

In Big education data, a variety of data is collected, stored and explored to unlock the value accrued from Big data. This section presents a literature review of previous works from three aspects: (1) Educational data collection; (2) Educational datasets; and (3) Educational data mining.

A. EDUCATIONAL DATA COLLECTION

Traditionally, educational researchers have been using methods such as surveys, interviews and classroom activities for data collection about student learning and experiences. Educational data can be collected at a rapid pace with the advance of online technologies (e.g. MOOC and LMS) which have the capability to track and collect a huge amount of educational data about learner experience. The Experience API (xAPI) [25] is an open data specification for data collection across learning tools. The authors in [43] use the xAPI standard to collect, track and store educational data retrieved from an e-learning environment called Kalboard 360. The tracked data is classified into three features (behavioral, demographic and academic background features). Another major source of educational data can be obtained from social media (e.g. blogs, online social networks, microblogs). It is challenging to collect social media data related to student learning experiences and behavior because of the variety and diversity of the language used. The authors in [37] performed data collection from Twitter using an educational account on a commercial social media monitoring tool.

B. EDUCATIONAL DATASETS

Educational datasets can be considered from two aspects [44]: (1) Datasets directly related to educational information containing educational resources, institutional data and educational indicators; and (2) Datasets from different domains which may be used in educational settings. Some examples of educational datasets are DBpedia [45], Freebase [46] and GeoNames [47]. The data in these educational datasets can be used for enriching the available educational content, discovery of new information which can help educational practices and connecting local datasets to the cloud. For a few examples, the authors in [48] used DBpedia to analyze the ranking of universities based on their structured information, and the authors in [49] used the categories provided by DBpedia to select the suitable categories for describing learning objects. Examples of datasets from different domains which may be used in educational settings include TEDTalks [50] which contains various conferences on a wide range of topics. Examples of other datasets could be from different domains and fields such as agriculture, medicine and tourism. Examples of datasets cited in the agriculture field are organic.edunet, Agris, AGROVOC, ASFA and JITA. Some examples of datasets cited in the medical field are PubMed and mEducator. PubMed is a service of the US National Library of Medicine which includes citations from MEDLINE and other scientific journals in life sciences. The mEducator Linked Educational Resources dataset is intended to provide educational resources in a linked data format, and are focused on the medical field, covering content ranging from traditional teaching to open learning, and experimental studies. Another project cited by several studies in the education domain is LinkedUp which have the objectives to collect and make available various types of data sources relevant for education, to provide a shared resource and to develop the community interested in the Web of Data for Education [51]. Other examples of university initiatives for linked datasets include the University of Southampton Open Data service, the Greek University Open Data, and the Linking Italian University Statistics Project.

C. EDUCATIONAL DATA MINING

Data mining techniques are increasingly gaining significance in the education sector and the outcomes from data mining techniques can provide invaluable support for decision making. The field of data mining in education is termed as Educational Data Mining (EDM). EDM is an emerging discipline that focuses on applying data mining tools and techniques to education related data. This section presents a literature review of the literature or survey papers for EDM and highlights their main contributions. A recent literature review or survey paper can be found in [11]. This review presents twenty years of data mining research in e-learning environments, from an educational perspective. The authors identified and classified challenges for research to improve student learner performances. Another literature review paper by [19] published in 2019 focused on EDM and learning analytics in higher education. The work in this literature review covered four main areas: (1) computer-supported learning analytics (CSLA) and the use of DM techniques to derive actionable information based on student interaction in LMS environments; (2) computer-supported predictive analytics (CSPA) and the use of EDM and LA to predict student performance and retention in courses based on assessment, engagement and domain knowledge in a learning activity; (3) computer-supported behavioral analytics (CSBA) and the use of DM techniques to identify student behavioral patterns and preferences when participating in online learning activities; and (4) computer-supported visualization analytics (CSVA) and the combination of information visualization techniques with advances in data mining and knowledge representation to offer a visual analysis of student behavior with respect to the learning activity.

Other review papers on EDM for education can be found in the works by [12]–[18], [52], [53]. Table 2 shows a summary of the various surveys which have been proposed for EDM. The table gives various details including the year, survey objectives, and remarks and comments. The authors in [15] surveyed the history and applications of data mining techniques in the educational field for traditional educational system, web-based educational system, intelligent tutoring system, and e-learning. The authors discussed concepts for EDM such as prediction, clustering, relationship mining, outlier detection, text mining, and social network analysis. In [12], the authors targeted to highlight the main data mining techniques applied in the e-learning environment and proposed three useful orientations for EDM research: (1) Orientation towards students and using EDM to recommend activities, resources and learning tasks to learners based on the tasks already accomplished by the learner and their successes; (2) Orientation towards educators and using EDM to obtain objective feedback for instruction, evaluate the structure of the course content and its effectiveness on the learning process; and (3) Orientation towards academics and administrators and using EDM to set parameters to improve site efficiency and adapt it to the behavior of users.

The authors in [13] presented a systematic review on EDM focusing on clustering algorithms and its applicability and usability in the context of EDM. The authors term this approach when applied to analyze datasets from educational systems as Educational Data Clustering (EDC). Different approaches for EDC were reviewed including 166 studies for e-learning and clustering, examination failure and clustering, intelligent tutor system and clustering, learning style and clustering, student modeling and clustering, student motivation and clustering, student profiling and clustering, etc. In [14], the authors performed a literature review focused on the different agents in the educational context as students, educators, researchers, institutions, and managers. The survey reviewed DM techniques applied to education, and models to provide updated information and improve institutional efficiency. The review of techniques included forecast performance modelling, undesired behaviour detection, monitoring support, recommendation planning and scheduling, and intelligent tutoring. Other review works on EDM can be found in [16]–[18]. The literature review paper of [16] discussed an explanation of the DM techniques in order of relevance, tendencies, and limitations faced by e-learning environments. In [17], the authors introduced a new perspective on the individualization and interaction between the educational actors and highlighted the trends and challenges of EDM from the perspectives of educational actors. In [18], the authors discussed the results of researches upon the behavior detection, personalization and student's performance evaluation obtained by DM techniques such as clustering, classification, and regression. In the work by [52], the authors focused on detecting the students' circumvention risks through predictive models and provide a custom recommendation to students by identifying their needs and learning disabilities. The objectives were to present a literature review of EDM focused on student's retention and evasion, recommendation systems and course administration. The work in [53] covered ubiquitous and pervasive data mining applied to education for fraud detection and identification of students that require special attention.

V. TECHNOLOGICAL ASPECTS FOR BIG EDUCATION DATA

In this section, some common platforms for Big data such as Hadoop, Spark and Samza will be discussed. Hadoop, Samza and Spark are currently the popular systems for Big data analysis. Hadoop is used for off-line and complex educational Big data processing, Samza is mainly used to solve the high data rate and large amounts for streaming education data processing, and Spark is often used for off-line rapid education Big data processing. The authors in [6] provided a general overview of Big data computing and discussed main characteristics such as data organization, decision-making, domain specific tools and platform tools. The authors illustrate the infrastructure that enables users to extract the maximum benefit from the large amounts of data available. In our context of Big data in education, this section aims to give a literature review for the Big data architectures or frameworks specially proposed for education. The architectures or frameworks on higher education setting is our focus. Specific software tools for data analytics/Big data which are increasingly being used in education will also be discussed.

A. BIG DATA PLATFORMS

Big data can be handled on different platforms. Hadoop and Spark are two commonly used platforms. In general, Apache Spark is used to manage massive amounts of data and to provide real-time analytics power.

1) HADOOP PLATFORM

Hadoop is an open source, distributed data processing distributed system infrastructure developed by the Apache Foundation. It enables distributed and parallel processing of large amount of data sets across clusters of many computers. It features low cost, high efficiency, high reliability, high scalability, and high fault tolerance. Hadoop consists of the HDFS distributed file system, MapReduce and several general-purpose tools.

MapReduce MapReduce is a paradigm of parallel programming across big datasets working with many computers (nodes). It supports the use of inexpensive computer clusters to perform distributed parallel computing on large datasets up to petabytes. The data can be in the form of structured or unstructured forms (e.g. weblog records, e-commerce click trails, binary or multi-line records). It is mainly composed of two functions: (1) Map function; and (2) Reduce function. The Map function is responsible for processing standardized data whereas the Reduce function mainly summarizes the results after the Map function.

HDFS is a distributed, scalable and portable filesystem for the Hadoop framework written in Java. HDFS stores large files (from gigabytes to terabytes) across many servers. HDFS provides unstructured data storage for Big data. HDFS is characterized by ‘‘write once read many times’’ and is very suitable for reading Big data. HDFS is a typical master-slave architecture. HDFS has the advantages of high fault tolerance and high scalability.

Hive is a data warehouse infrastructure built on top of Hadoop which provides summarization of data, query and analysis. Hive supports analysis of big datasets stored in HDFS, Amazon S3 file system etc. It provides an SQL –like language called HiveQL, supporting indexes.

NoSQL: is a database system providing a mechanism for storage and retrieval of data with less constrained than traditional SQL (relational) databases.

Hadoop Common provides java libraries and utilities which are required by other Hadoop modules.

Mahout: Mahout is an open source machine learning and data mining algorithms sets based on Hadoop which has implemented many machine learning and data mining algorithms.

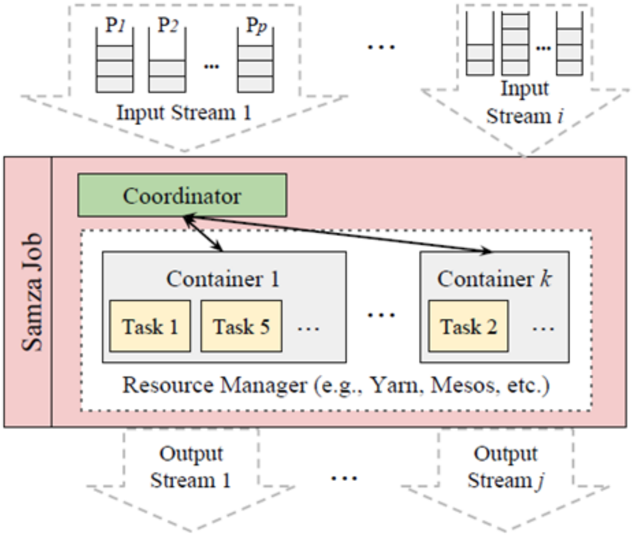

FIGURE 2. Samza technology core architecture [56].

Other Business Intelligence (BI) Tools Although an increase Big data technology is huge, it doesn’t mean the end of classical BI tools like Cognos, QlikView, SPSS and so on. The trend is that BI tools would be able to work with new Big Data technologies side by side.

Data Storage: NoSQL databases are inherently schema less and highly scalable. These databases support frameworks like MapReduce, Dryad etc. for the parallel processing of large amounts of data. The paper by [54] investigated educational technology for Big data analysis and the exploration of the development trend for online education. The authors gathered data, attached importance to the basic function and value of education data, and explored the education technology that matches the Big data analysis. The work by [55] discussed the relationship between Big data and cloud computing, Big data storage systems and Apache Hadoop technology.

2) SPARK PLATFORM

Apache Spark is a distributed computing framework like MapReduce but maintains data in Resilient Distributed Dataset (RDD). It is useful for algorithms that perform iterative operations and data flow processing. Spark provides Shark, an interactive query analyzer, Bagel, a high-volume graph processing and analyzer, Spark Streaming, a real-time analyzer, and Mllib, a machine learning library.

3) SAMZA PLATFORM

Samza is a distributed stream processing framework for real-time data processing. In Samza, the data stream is partitioned, and each partition is given a specific ID or offset. Samza places the storage and processing on the same machine and does not load additional memory while maintaining processing efficiency and providing a framework for a flexible pluggable API. Fig. 2 shows the Samza technology core architecture [56].

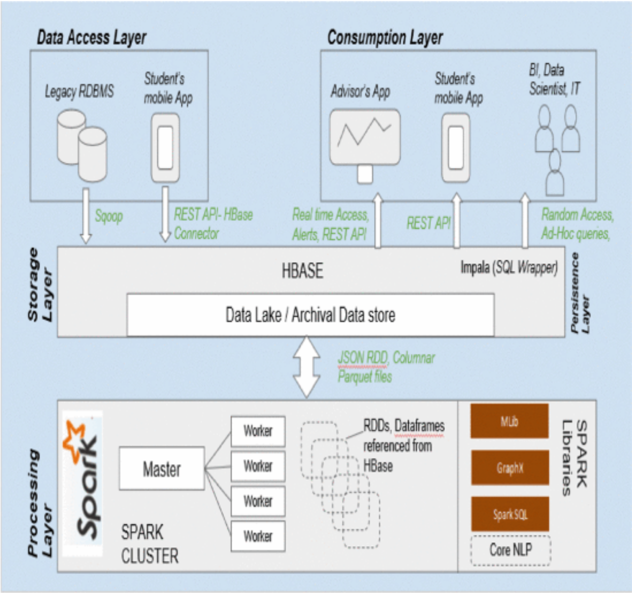

FIGURE 3. Distributed architecture for big education data [1].

B. FRAMEWORKS AND ARCHITECTURES FOR BIG EDUCATION DATA

This section discusses several frameworks and architectures for Big education data. The authors in [1] proposed a distributed architecture for the information processing of Big education data. The authors use this architecture to predict student performance with and without sentiment analytics. Fig. 3 shows their proposed architecture which consists of three layers: (1) Data Access Layer; (2) Data Storage Layer; and (3) Data Processing Layer. The Data Access Layer comprises of all the data sources the processing engine require for the information processing such as student logs, student records and historical data), and a student mobile application which can generate data based on a student's activity. The data sources are connected to the Storage Layer (HBASE) using the Sqoop and REST API-HBase Connector. The second layer is the Data Storage Layer which comprises of HBase and the HDFS distributed storage. The third layer is the Processing Layer which performs the sentiment and predictive analytics. This layer uses the Spark cluster. In this layer, the features were transformed to the Spark Resilient Distributed Data (RDD) formats to perform the predictive analytics. The predictive modeling procedures were performed via a process of ensemble modeling.

The authors in [2] proposed an architecture termed *Concept Definition for Big Data Architecture in the Education System*. Their architecture consists of five layers: (1) Data Sources; (2) Big Data Processing; (3) Data Warehouse; (4) Data Mining Tools; and (5) Reporting. In the Data Sources layer, the data can be stored in traditional SQL databases (e.g. classical relational data) or NoSQL databases (e.g. data from social networks). The Big Data Processing layer uses Apache Hadoop to process the huge amount of data from the earlier layers. The third layer is the Data Warehouse layer which is technology and vendor independent for creating the data cubes. The Data Mining layer uses tools from IBM SPSS or SAS. The highest layer (Reporting Layer) performs the creation of useful analysis from the obtained data for different types of users (e.g. teachers, administrators, other stakeholders for the university). The Cognos software application from IBM could be used for the reporting functionality.

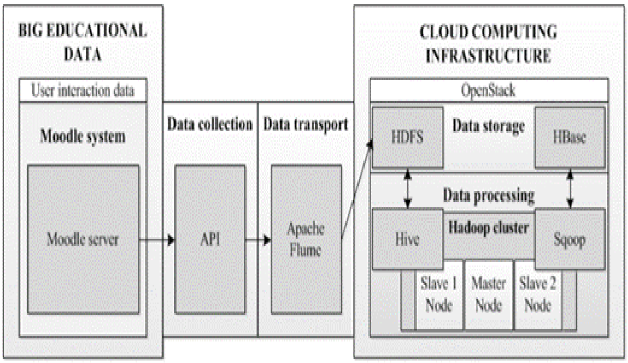

FIGURE 4. Processing of big educational data in the cloud [3].

The authors in [3] proposed an architecture to analyze educational data from the Moodle system in the cloud using Apache Hadoop. Their cloud-based architecture consists of four stages: (1) Big Educational Data; (2) Data Collection; (3) Data Transport; and (4) Cloud Computing Infrastructure. The Big educational data are collected through the API or other interfaces and transported to the data storage with the use of the most suitable platform, tool or service. The data storage and data processing are performed in the cloud. The data-intensive computing framework is applied to analyze massive amounts of data to reveal the valuable information. The main contribution of this paper is the newly proposed model approach for processing big educational data generated from the Moodle system, which was also implemented and validated as an experimental architecture as shown in Fig. 4. The architecture was constructed based on open-source platforms, tools and services. The API was used to limit programming only to the computational tasks and data transfer from the Moodle system to the cloud. The experimental implementation of the proposed model approach was performed with the use of the following platforms: Apache Flume, Apache Hadoop and Hadoop Distributed File System (HDFS), Apache HBase, Apache Hive, Apache Sqoop and OpenStack.

The authors in [4] presented a Big data architecture for education using Spark. As shown in Fig. 5, the various data are delivered in HDFS according to each attribute. The structured data is transferred from the RDBMS to HDFS using SQL-to-Hadoop (Sqoop). Among these collected data, lecture data is an important item, and the FP-Growth algorithm is performed using MLlib, a spark machine learning library. The resulting data can be used to identify patterns of lecture

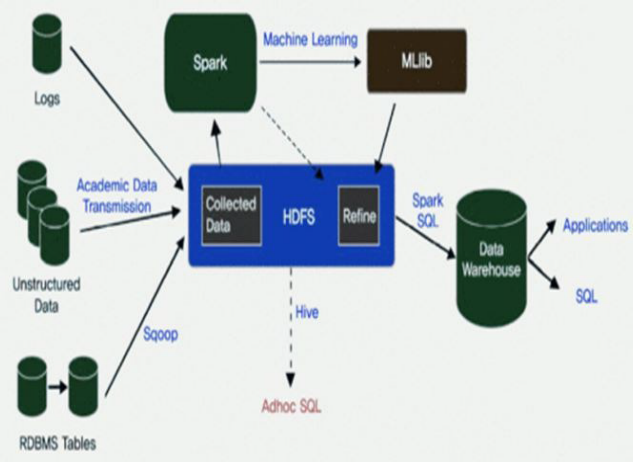

FIGURE 5. Educational big data architecture using Spark [4].

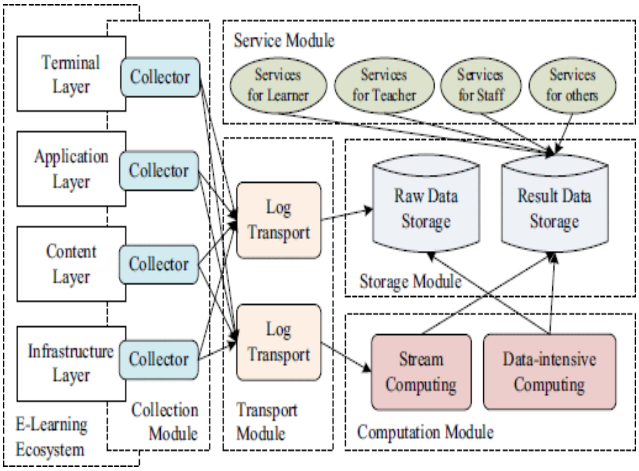

FIGURE 6. E-learning big data ecosystem [5].

data that students have taken for the year and semester. These patterns include pattern information for students' preferred lectures, and based on this, a recommendation system was implemented that recommends lectures to students. In addition, by using the data collected from the sensor information of the classroom attendance and the dormitory entrance information, it is possible to determine the population of the students, predict the density of the population and to control the temperature of the classroom and buildings by pattern analysis.

The authors in [5] proposed an architecture for an E-Learning Big Data Ecosystem. It is composed of five modules as shown in Fig. 6: (1) Collection Module; (2) Transport Module; (3) Storage Module; (4) Computation Module; and (5) Service Module. The Collection Module contains collectors distributed in each layer. Each collector records the log data produced by different objects and normalizes the collected data. The Transport Module transfers the collected log data to where the data is required. The Storage Module includes two categories of storage systems. The first storage system is the Raw Data Storage system which stores historical data for future data mining and analyzing. The second storage system is the Result Data Storage system which has the ability of rapid access to data to provide high I/O performance for stream computing in the Computation Module and the Service Module. The Computation Module contains two computing frameworks. The first is the data intensive computing framework which is applied to analyze massive raw data to dig out valuable information, and the second is the stream computing framework which is applied to deal with every coming data in real-time. The computing frameworks are required to support parallel computing to guarantee low latency of the analyzing process and improve computational efficiency. The Service Module reads the needed data from the Result Data Storage system for all objects and roles in each layer of the e-learning ecosystem.

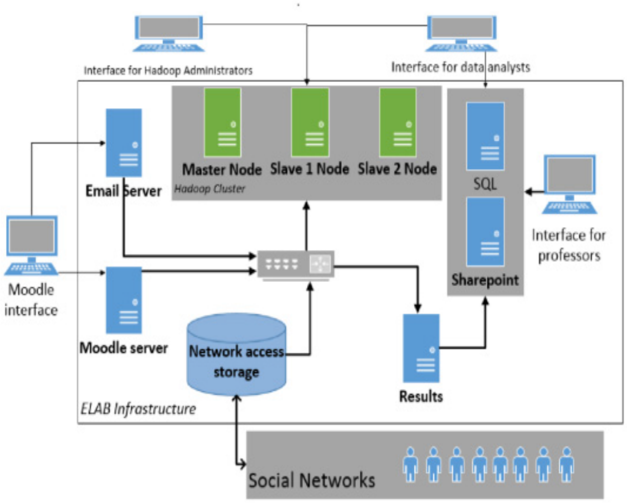

FIGURE 7. Big data Hadoop infrastructure for education [6].

The authors in [6] proposed a Big data infrastructure deployed as a Hadoop platform in order to improve the education process. The platform is integrated with the learning management system (LMS) Moodle platform. The platform is deployed within the e-learning infrastructure of a laboratory. Fig. 7 shows the implemented Hadoop e-learning infrastructure. The Hadoop cluster contains three nodes (Master node, Slave 1 node, and Slave 2 node). The Hadoop cluster is also connected to the Email server, Moodle server. Network data storage and Sharepoint cluster through the Results server. The cluster communicates with other components using the TCP/IP protocol and all data is transferred through the Ethernet infrastructure.

The authors in [7] proposed an architecture based on Apache's Hadoop open source distributed Big data computing architecture. It is used to process the Big data of Holland vocational interest theory. The core module is divided into two parts: (1) Hadoop Distributed File System (HDFS); and (2) Hadoop Parallel Programming Framework (MapReduce). The overall architecture of the system is composed of three layers: (1) Data Layer; (2) Logic Layer; and (3) Presentation Layer. The Data Layer supplies the basic data supporting for the entire system and stores the mass data of student behavior data including teaching, education management, scientific research, campus life and so on. The Logic Layer is the core part of the whole system, which is the value of the data mining. The Presentation Layer provides a visual interface for users. The graphical data analysis interface can help users to perform the Holland analysis, curriculum optimization and student employment decision. Other works on frameworks and platforms for Big education data can be found in [8]–[10]. The work in [8] used the Hadoop platform to conduct parallel mining of educational literature on Big data. The paper has analyzed the main function of text mining technology, and combined Canopy and the k-means algorithm to analyze and research the educational Big data literature. The authors in [9] presented a framework for a Big data education system based on Hadoop. They examined the MapReduce system for the education system and the huge volumes of data were stored in HDFS. The authors in [10] provided a comparison on the Hadoop, Spark and Samza platforms, and presented an architecture of Spark for education.

VI. DATA ANALYTICS FOR BIG EDUCATION DATA

This section gives comprehensive discussions for data analytics for Big education data from two areas: (1) Predictive analytics; and (2) Learning analytics. A brief literature review of some emerging trends and opportunities in applications of Big data in educational data mining and learning analytics can be found in [57] and [58].

A. PREDICTIVE ANALYTICS (PA)

The prediction of how well a student or a group will perform on a learning task is one of the most popular and useful applications of educational predictive analytics. It can also be used to identify at-risk students who are likely to fail. However, there is a challenging problem to solve due to the large number of circumstances that can impact student performance, such as socioeconomic status, cultural background, demographic characteristics and psychological profile. This section gives discussions for predictive analytics from three application areas: (1) Student performance; (2) Dropout prediction and academic early warning systems; and (3) Courses selection.

1) STUDENT PERFORMANCE PREDICTION

The authors in [1] provides a discussion on Big data, learning analytics and use of natural language processing (NLP) in higher education. They proposed an integrated analytics model with predictive analytics for student performance on their Big data architecture with data access, storage and processing layers. The architecture has been discussed in Section V. Their analytics model utilizes different types of data to predict student performance and support student progress. The authors incorporate the usage of sentiment analysis in their predictive analytics to and employ a distributed technology system capable of supporting academic authorities and advisors at educational institutions in making decisions. Their experiment results showed that the features derived from unstructured data gave a 10% improvement in the accuracy of results compare with the traditional single predictive model. The authors in [59] proposed an approach using predictive analytics for e-learning with the Hadoop Big data platform. Their work used the decision tree classification approach (C4.5) in a Hadoop framework to predict student performance. The C4.5 algorithm was proposed because: (1) It is able to handle both discrete attributes, and continuous attributes; (2) It can process partially complete training data sets with values not present; (3) Pruning can be done while constructing the trees to prevent the over-fitting problem. The work by [60] proposed a two-stage model, supported by data mining techniques that uses the information available at the end of the first year of students' academic career (path) to predict their overall academic performance. This study proposed to segment students based on the evidence of failure or high performance at the beginning of the degree program, and the students' performance levels predicted by the model. A data set of 2459 students spanning the years from 2003 to 2015 from a European Engineering School of a public research University was used to validate the proposed methodology. The empirical results demonstrated the ability of the proposed model to predict the students' performance level with an accuracy above 95%.

The ASSISTment [61] system designed by Worcester Polytechnic Institute and Carnegie Mellon University can tutor students and assess the student learning at the same time. This system targets the problem that instructors wish to do assisting and assessing at the same time in class. The system gives assessment results by predicting the student's performance on standard test given by official assessment system such as MCAS (The Massachusetts Comprehensive Assessment System). It collects the student's reaction information (such as accuracy, speed, the number of hints required and performance on sub-steps) and predicts the student's performance based on the correlation model trained by past data of past months and years. Since the students work on the system every week, the ASSISTment system can keep updating the value of metrics and provide increasingly accurate predictions. The authors in [62] developed a predictive model to forecast the student performance in higher level modules based on the contextual factors. The authors analyzed data from 1037 students across various specializations, with different mode of study, age group, gender and different sponsors. The Rapid Miner open source tool for predictive analytics and visualization was chosen for the study. The outcome of the work showcased that negative correlation exists between age and the academic performance, whereas positive correlation exists between lower level and higher-level modules.

Other examples of predictive analytics for student performance can be found in [63]–[70]. The authors in [63] used student information like attendance, class test, seminar and assignment marks collected from the student management system to predict the performance at the end of the semester. This paper investigated the accuracy of decision tree techniques for predicting student performance. The work in [64] analyzed live video streaming and the students online learning behaviors and their performance in their courses. The student participation and login frequency, as well as the number of chat messages and questions that they submitted to their instructors were analyzed together with the student's final grades. The results of the study showed a considerable variability in students' questions and chat messages and revealed that combining EDM with traditional statistical analysis provides a strong and coherent analytical framework capable of enabling a deeper and richer understanding of students learning behaviors and experience. The authors in [65] explored the use of predictive modeling methods for identifying students in virtual learning environments (VLE) who will benefit most from tutor interventions. The methods discussed included decision-tree classification, support vector machine (SVM), general unary hypotheses automaton (GUHA), Bayesian networks, and linear and logistic regression. The methods were trialed through building and testing predictive models using data from several Open University (OU) modules. This work highlighted the importance of understanding how a student's pattern of behavior changes during the course. The authors commented on two findings: (1) VLE activity is a useful data source to include for predicting student outcome but should not be viewed as an absolute measure of engagement but rather with reference to a student's own past behavior; and (2) Feature selection has a big impact on the reliability of a model generated from the data regardless of which model type is chosen.

The work in [66] demonstrated how web usage mining can be applied in e-learning systems to predict the marks that university students will obtain in the final exam of a course. In this work, the authors developed a specific Moodle mining tool oriented and compared the performance of different data mining techniques for classifying students. Several well-known classification methods were used such as statistical methods, decision trees, rule and fuzzy rule induction methods, and neural networks. The authors carried out several experiments using available and filtered data to try to obtain more accuracy. The authors in [67] used predictive analytics to identify the factors influencing the performance of students in final examinations and found a suitable data mining algorithm to predict the grade of students. The authors designed a neural network (multilayer perceptron) tool using the .NET framework to predict the grade of the student when given the various parameters as input and achieved an accuracy of 72% which showed the potential efficiency of the MLP algorithm. The obtained results from hypothesis testing showed that the type of school did not influence student performance and on the other hand, the parents' occupation played a major role in predicting grades. The work in [68] proposed an approach to predict student performance through genetic programming. The authors used activity theory derived participation indicators as inputs into a Genetic Programming (GP) model to develop a student performance prediction model. Their GP model was able to build a prediction model without assuming any a priori structure of functions. The proposed GP model also provided instructors with individualized suggestions to students in any performance state (at-risk, just survive, average or good) as well as increasing students' awareness.

The authors in [69] proposed an educational data mining (EDM) case study based on the data collected from learning management system (LMS) of e-learning center and electronic education system of Iran University of Science and Technology (IUST). The authors implemented a model to predict the GPA of graduated students. To achieve goals, a common methodology of data mining was utilized which is called CRISP. Our results show that there can be confident models for predicting educational attributes. The work in [70] also used data mining as a predictive tool for performance improvement of engineering students. The authors applied the C4.5, ID3 and CART decision tree algorithms on engineering student data to predict their performance in the final exam. The authors showed that the outcome of the decision tree classifiers predicted the number of students who are likely to pass, fail or promoted to next year. Their results provided steps to improve the performance of the students who were predicted to fail or promoted. The comparative analysis of the results also showed that the prediction has helped the weaker students to improve and brought out better outcomes in the result.

2) DROPOUT PREDICTION AND ACADEMIC EARLY WARNING SYSTEMS

One of the biggest challenges every institution face is how to improve student retention and reduce attrition. There could be several reasons for student attrition including academic issues (inadequate preparation, student disinterest with content or delivery method); motivational issues (low level of commitment to the institution, perceived irrelevance of the institution's experience); psychosocial issues (social factors, emotional issues); and financial issues (inability to afford fees, perception that cost outweighs benefits) [71]. Two emerging areas to improve student retention and reduce attrition are (1) Dropout prediction; and (2) Development of academic early warning systems. Dropout prediction is one of the major research topics in learning analytics (LA) for Big education data. The prediction of dropout is very useful to instructors and to be able to identify how likely a student would drop out during the course. The instructor can make some adjustments during the teaching process to mitigate and reduce the likelihood (e.g. send email reminders or give positive feedback to students who have been identified to be very likely to drop out during the course).

Some examples of LA for dropout prediction can be found in the works by [72]–[80]. The authors in [72] investigated dropout prediction in massive open online courses (MOOC). The objective was to predict from the student behavior log data the likelihood of students dropping out from the MOOC in the next ten days. In this work, the authors collected 39 courses data from the XuetangX platform which is one of the largest online learning platforms in China. The authors used four supervised classification models (SVM, logistic regression, random forest and gradient boosting decision tree (GBDT)) to perform the dropout prediction task and achieved the highest classification accuracy of 88% accuracy with the GBDT. The work in [73] used machine learning (ML) techniques to demonstrate that categorizing student performance data and exercise sets were adequate parameters for identifying possible dropouts during a course. The authors used experimental data from a computer science course and showed that their ML techniques could provide automatic detection of student dropouts during the second week of the eight-week courses.

The work in [74] utilized education data mining to analyze the factors affecting student academic performance which contributed towards the student failure and dropout. The authors showed that their techniques enabled the identification of weak students shown to have poor performance. The authors in [75] used learning analytics to manage dropout rates based on a set of pedagogical actions in distance education courses and reported an average of 87% prediction accuracy and an average reduction of 11% in dropout rates. Other works for dropout prediction can be found in [76]–[80]. The authors in [77] conducted experiments using a dataset of 419 students to determine the best predictors of dropout at different stages in a course. The authors in [77] extracted features from student behavior from completed curriculum and applied machine learning algorithms to predict the dropout rate. The authors in [78] used data mining algorithms to predict student failure from high dimensional and imbalanced behavior data. A second emerging area in LA for Big education data is the development of academic early warning systems (AEWS). The objective of an AEWS is to discover and identify existing and potential academic problems of students in the early stages of education and inform students so that remedial actions can be taken to mitigate the risks. The authors in [81] proposed an AEWS based on Big education data collected from different departments of the university such as the academic affairs, library and other departments. The authors used principal component analysis (PCA) to locate the key predictors and utilized three machine learning algorithms to train and test their classifiers from their sample data. Their results showed that the naïve Bayesian algorithm gave the best accuracy rate of 86% for three-semester data and 85.4% for one-semester data.

3) COURSES SELECTION

This section focuses on the articles or works where learning analytics is used as a tool for courses selection. The authors in [82] proposed a system termed as Degree Compass to be used by students who are not familiar with navigating their way through a degree program. The Degree Compass system uses data from hundreds of thousands of past students with the data of a particular student (course grades, standardized test scores, college transcript grades, etc.) to recommend courses to students that is most likely to achieve the best grade and which also fits with the program of study of the student. The system has been shown to be able to correctly distinguish if the student will get either an ABC grade or a DF grade with 92% accuracy. The authors in [83] proposed an approach for Big data analytics for predicting academic course preference using Hadoop and MapReduce. In their work, they derived preferable courses for pursuing training for students based on course combinations. The input dataset collected from students is split into various clusters and provided to the mapper that maps data to the output which are represented as <key, value > pairs. The output obtained from the mapper are then combined in the combiner and then sent to the reducer. The authors in [84] developed educational models to predict how learning materials might be designed to fit the knowledge of the student. Their approach used educational data mining to develop educational models to predict how learning materials might be designed to fit the knowledge of the student.

B. LEARNING ANALYTICS

Learning Analytics (LA) is the collection and analysis of usage data associated with student learning. This section gives discussions for LA from five areas: (1) Collaborative and interactive Learning; (2) Behavior learning; (3) Personalized learning; (4) Social network analytics; and (5) Learning and assessment analytics.

1) COLLABORATIVE & INTERACTIVE LEARNING

Collaborative analytics are commonly used to deal with issues related to providing instructional strategies that supports and enhances the collaboration process among students who work together in small groups. A collaborative learning environment (CLE) aims to improve continuous and reciprocal student-educator interaction, cooperation towards knowledge construction, and knowledge and experience exchange to reach common goals. The work in [85] presented an empirical case study to investigate the impact of collaborative learning patterns on student achievements with educational data captured from a CLE platform. The authors analyzed the progress time series reflecting students' contributions to an assignment to investigate different styles of collaborations. By comparing the collaborative learning patterns of the same groups in completing different assignments, the authors explored the pattern impact on the grades received as a result of teacher assessments of these assignments and identified the characteristic patterns that lead to better learning outcomes either in terms of quality or efficiency. The authors showed that continuous focus, self-reflection, live collaboration, and even distribution of workload and contributions were more likely to lead to more refined and coherent assignments, and consequently achieve better marks. A different approach was taken by the authors in [86] which proposed using student interaction to measure the effectiveness of collaboration in virtual learning environments (VLE). In this work, the user activity logs from the learning platform were used as the main tool for inferring learners' activities to fit certain behaviors and preferences. The work by [87] examined the effects of learning analytics as supporting tools for instructors to guide cooperating groups. Other examples of papers on collaborative and interactive learning for LA can be found in the works by [88]–[91].

2) BEHAVIOUR LEARNING

The concept of behavior learning is important to understand student learning and evaluating student performance. The authors in [92] proposed searching for student behavioral patterns while accessing and browsing educational resources. In this work, the authors extracted behavioral patterns related to the student interactions with the educational media. Their results demonstrated the usefulness of student perception and identified the trends regarding the use of educational media for learning. The authors in [93] developed an evaluation system for student learning through the factors analysis that influences their behavior during the media usage. The goal was to improve the evaluation method in order to improve the students' behavior in relation to use of the learning media. To evaluate the level of student's learning, the decision tree technique was used. The authors in [94] developed a system to explore and visualize generated data in virtual learning environments and analyzed these data using web-mining and statistical techniques to extract behavior patterns of the student. The authors in [95] grouped and analyzed access data in order to recognize behavior patterns (e.g. identify whether the instructions were inadequate or insufficient, or to identify visibility problems in the content posted) in order to review and organize the educational content. The authors in [96] presented a framework for analyzing student activity data in open-ended learning environments (OELE) that integrates model-driven behavior characterization and data-driven pattern discovery. The model-driven approach used linked task and strategy models to provide more precise interpretation of student activity sequences as learning and problem-solving strategies while the pattern mining approach enables the identification of new variations of strategies and of gaps in the coverage of the current strategy model. Other examples of papers on behavior learning can be found in the works by [97], [98].

3) PERSONALIZED LEARNING

Personalized learning is aimed at customizing the learning journey of a student to maximize his/her learning potential and hence fulfill the goal of education and career with satisfaction and accomplishment. With the help of Big data technologies, learning can be made increasingly personalized, and instructors can watch learners and track which areas within a program of study they find challenging and spend most of their time, the learning materials they revisit often, the sections they recommend to their peers, the learning styles they prefer, and the time of day they learn better [99]. With the emergence of various learning strategies such as micro-learning, multimedia learning and flipped classroom, learning personalization has been recognized as an effective and adaptable interface between the student and the knowledge to allow effective learning and knowledge transfer. For example, in micro-learning, information is delivered in small portions that are easy to learn effectively [100] and content can be delivered according to a tailored knowledge composition patterns that are best retained by individual students. Personalized learning has been advocated as an effective approach that could be applied at different stages of the curriculum to ensure deep learning and leaves students with knowledge absorbed quicker and retained longer.

4) SOCIAL LEARNING AND NETWORK-BASED ANALYTICS

Social and networked-based learning and analytics benefit from the utilization of technology to establish connections between students, instructors, communities and resources [101]. The use of EDM and LA for social networks analysis has been reported to be associated with student learning and building knowledge in social and cultural settings to discover patterns of collaboration, assessment and communications. The work by [102] showed that by collecting data about user behavior, LA could be useful for providing recommendations about learning resources and activities. The work by [103] showed that mining students' online social interaction was important for recommending appropriate learning partners in a web-based cooperative learning environment. Another work for EDM and LA to aid educational decision makers by providing the environment to share and collaborate with other team members to take the appropriate actions for a given learning task can be found in [104].

5) LEARNING & ASSESSMENT ANALYTICS USING EXPERIENCE API

The Experience API (xAPI) standard is a specification for learning technologies which can be used for data collection describing the wide range of experiences of the learner in the context of formal learning, informal learning and social learning [105]. The authors in [109] gave two classifications for research works using the xAPI specification in the context for learning analytics: (1) The first category deals with the deficiencies of xAPI specification such as limitations of learning interactions and inconsistency of learning behaviors across platforms in addressing specific issues related to the learning context; and (2) The second category deals with tracking and analyzing the learning experience using the xAPI specification. The work by [43] used the xAPI standard to track educational data from an e-learning environment called Kalboard 360. The tracked data is classified into behavioral, demographic and academic background features and three data mining techniques (ANN, naive Bayes and decision tree classifier) were employed to evaluate the impact of such features on student performance. The experimental results showed that there was a strong relationship between learner behaviors and their academic achievement. The authors in [107] proposed a 3D design activity stream for STEM education based on xAPI. The xAPI can describe learner experiences as active statements with eight attributes (UUID, ACTOR, VERB, OBJECT, RESULT, CONTEXT, TIMESTAMP and VERSION). For example, the specification <ACTOR, VERB,

OBJECT, CONTEXT > composes a simple activity flow. Experiments were carried out at the Li Jun School in China. The authors collected more than 22,000 data elements and showed that their *x*API could completely record the learning paths of students. Their results also showed that students had different operating habits and learning paths which provided the basis for the evaluation of students' spatial thinking ability and engineering design skill in the interactive learning environment. The authors in [108] discussed some experiences and learnt lessons from implementing *x*API for projects in the Netherlands. The authors remarked on the need for a centralized approach for data collection to get a complete picture of student behavior which may be stored on many heterogeneous IT systems. Furthermore, the *x*API recipes need to be seen in their infrastructural context. An ETL (Extract Transform Load) layer with communal best practices encoded in the transforms and applied across the higher education sector can enforce the authoritative standard and decrease the overall costs.

The authors in [106] discussed a case study to show the suitability of using *x*API (Tin Can API) for self-regulated learning (SRL). The authors proposed an extension of *x*API for recording SRL-related actions termed as *x*API-SRL. Their monitoring system had several steps: (1) Author – filter statements from the selected author; (2) SRL – filter SRL related actions; (3) Time – select time window and organize records time wise; (4) Object – filter or organize statements attending to the object; (5) Grouping and analysis – analyze groups of statements attending to how they relate to each other. A recent work by [109] explored the use of *x*API in learning analytics for MOOC environments which generated big assessment data (Big data) given the massive number of courses proposed and the high number of learners enrolled. These assessment data must be tracked, processed and analyzed as the learning data. The authors in [110] commented that assessment analytics has the potential to make valuable contributions to the field of learning analytics by extending its scope and increasing its usefulness. The authors also state that the role that assessment analytics could play in the learning process is significant and yet it is underdeveloped and underexplored.

C. RECOMMENDATION SYSTEMS

A recommendation system or recommender is an information filtering system that seeks to predict the rating or preference a user would give to an item. These systems have been very helpful in applications such as e-commerce (e.g. Amazon), entertainment (e.g. Netflix, YouTube and Spotify), service industries, and social media platforms (e.g. Twitter and Facebook). Recently, recommender systems have gained popularity in the education sector to generate various kinds of recommendations for learning institutions, instructors and students. This sub-section explores recommendation systems for Big data in education. The various recommendation techniques can be broadly categorized into four types [111]: (1) Collaborative-based filtering; (2) Content-based filtering; (3) Knowledge-based systems; and (4) Hybrid-based systems. In collaborative-based filtering systems, an item will be recommended to the user based on the preference of other similar users for the same item. The sets of users which have the strongest correlation in the past will be identified as nearest neighbors, and the score of the new items will be predicted based upon the scores of its nearest neighbors. The correlation or log-likelihood ratio measures can be used to identify preferred items for the user. Content-based filtering recommender systems utilize a series of discrete and pre-tagged characteristics of an item in order to recommend additional items which have similar properties. Content-based recommendation systems find out items of interest for users by analyzing item descriptions. These systems generate lists of item profiles for the users based on the data provided by users. It uses two metrics called term-frequency (TF) and inverse document frequency (IDF). The TF determines how many times the item has occurred in a document whereas the IDF identifies the importance of the item. The product of TF×IDF is used to identify the importance of the item. Knowledge-based recommendation systems are based upon the knowledge of a user's need for an item and can therefore reason about the relationship between a need and a possible recommendation. The knowledge about the user needs, preferences, etc. are used to perform the recommendation. Current recommender systems typically combine one or more approaches into a hybrid recommendation system to improve the recommendation accuracy. Examples of recommendation systems for educational data can be found in [112]–[119].

For specific course recommendation of MOCC, some approaches such as collaborative filtering, content-based filtering and hybrid recommendation systems can be found in [113]–[115], [116]. The authors in [113] proposed a systematic methodology for recommending personalized courses and considering the sequence of learning curriculum. In their system, they considered a measurable context space with Lipschitz condition, where space is divided into many subspaces to represent different types of students. The course clusters are defined to capture the prerequisite dependencies among courses. Their dataset is composed of three parts: (1) Data of courses; (2) Context information of the students; and (3) Feedback reward records. The course data was obtained from the biggest MOOC platform in China called 'iCourse' which contains nearly all the Chinese online courses. The context information was collected from 4939 anonymized students in Huazhong University of Science and Technology and Central China Normal University (∼20,000 learning records). The reward records are the scores of courses and the degree of satisfaction. The authors in [114] proposed a Big data solution on Hadoop platform for recommendation of pedagogical documents that meet the identified needs of the learner. This system will be established by using Big data as a tool to analyze the performance and skill level of students individually and then create personalized learning experiences that fit into their specific learning paths. The authors used a semantic approach which recommends learning objects by comparing the textual contents of resources that form a corpus of pedagogical documents and proposed an algorithm for similarity measurement between the document viewed by the learner and the documents of corpus of pedagogical documents available in order to select from those which are most similar to the viewed document. Their work was implemented and tested on the Hadoop Big data platform. For the implementation of the recommendation algorithm, modules were coded in Python using scikit-learn and NLTK python packages. For parallelization, MapReduce was leveraged to process the data stored in Google File System (GFS). The authors in [115] also designed and implemented a personalized recommendation system on Big data platform. Their system can help people to automatically excavate interesting and valuable information from target data. A personalized education resource recommendation system which can handle Big data is studied and implemented. The results showed that the personalized recommendation system of educational resources based on Big data has been put into use in a university network and achieved the expected design goal. This system, combining the discipline classification tree and the recommended structure, provides the resilient processing ability with the increase of data and the personalized recommendation function based on the security, high efficiency and real-time of Big data. It provides effective help for the students and teachers to make use of the valuable teaching resources. However, when they evaluated their recommendation algorithm, the MovieLens dataset (not educational data) was used to verify the performance.

Other educational recommendation systems can be found in [116]–[119]. The authors in [116] built a personalized English learning recommender system for students to set basic score of lessons. The collaborative filtering technique and content-based method was used. Another author [117] developed a recommender system for predicting student performance. Their approach mapped educational data to user/item. The matrix factorization technique was used to generate the recommendation and logistic regression to validate their approach. An automated recommender system for course selection can be found in [118]. The collaborative recommendation technique was used to recommend elective courses to students by using association rule mining to generate course association rules. The authors in [119] built a semantic educational recommender system in formal e-learning scenarios. They used a conceptual approach which can be used as personalized recommender in e-learning scenarios in their work. Other examples of earlier works on recommendation systems for e-learning can be found in [120]–[126]. The Recommendation Agent for e-learning systems is one of the first collaborative filtering educational recommendation systems that have been established [120].

D. GRAPH ANALYTICS

Graph analytics can be used to determine the strength and direction of relationships between objects in a graph. This section discusses some research works to address challenges of Big data from online education data using graph analysis. Some examples of using graph-based analytics and machine learning to address challenges and opportunities for education can be found in the works by [127]–[129]. The authors in [127] used observed prerequisite relations among courses to learn a directed universal concept graph and used the induced graph to predict unobserved prerequisite relations among a broader range of courses. This is particularly useful to infer prerequisite relations among courses from different providers e.g. universities, MOOC, etc. The authors proposed a new framework called Concept Graph Learning (CGL) for inference within and across two graphs at the course level and at the induced concept level. The explicit learning of the directed graph for universal concepts is the key part of the framework. Once the concept graph is learned, it could be used to predict unobserved prerequisite relations among different courses including those not in the training set and from multiple sources. Their experiments showed promising results for cross-universities setting. The universal transferability is particularly desirable in MOOC environments where courses are offered by different universities and instructors.

The authors in [128] addressed the graph analysis problem in multi-source relational learning for educational data. When the numbers of nodes in multiple graphs are large, the labeled training instances are extremely sparse. Existing methods such as tensor factorization or tensor kernel machines do not work well because of the lack of convex formulation for the optimization, the poor scalability of the algorithms in handling combinatorial numbers of tuples and the non-transductive nature of the learning methods which limits their ability to leverage unlabeled data in training. The authors proposed a Cross-graph Relational Learning (CGRL) approach for predicting the strengths or labels of multi-relational tuples of heterogeneous object types. They formulated the CGRL as a convex optimization problem which enable transductive learning using both labeled and unlabeled tuples and proposed a scalable algorithm that guarantees the optimal solution and enjoys a linear time complexity with respect to the sizes of input graphs. The authors conducted the experiments on 34,340 DBLP publication records in the domain of Artificial Intelligence. Tuples in the form of (Author, Paper, Venue) were extracted from the publication records leading to 15,514 tuples (cross-graph interactions) after preprocessing. The authors showed that their proposed method successfully scaled to the large cross-graph inference problem, and outperformed other representative approaches significantly. A recent work on graph analytics by [129] presented the early detection prediction of learning outcomes in online short course via learner behaviors. Through evaluation on data captured from three two-week courses hosted through delivery platforms, the authors made three key observations: (1) Behavioral data contains signals predictive of learning outcomes in short-courses (with classifiers achieving AUCs ≥ 0.8 after the two weeks); (2) Early detection is possible within the first week (AUCs ≥ 0.7 with the first week of data); and (3) Content features have an ‘‘earliest’’ detection capability (with higher AUC in the first few days), while the SLN features become the more predictive set over time as the network matures. They also discuss how their method can generate behavioral analytics for instructors.

E. VISUAL ANALYTICS

Visual analytics (VA) focuses on analytical reasoning facilitated by interactive visual interfaces and scientific visualization. This section gives some review and discussions and applications of VA in Big education data. The authors in [130] presented a systematic review of the emerging field for visual learning analytics of educational data. The authors found that: (1) Few works have been done to bring visual learning analytics tools into classroom settings; (2) Few studies have considered the background information from the students such as demographics or prior performance; (3) Traditional statistical visualization techniques such as bar plots and scatter plots are still commonly used in learning analytics contexts; and (4) While some studies employ sophisticated visualizations, there is a lack of studies that employ sophisticated visualizations and engage deeply with educational theories. Two other studies for visual data mining can be found in [131] and [132]. The use of VA methods can help turn the features of education into a visible type of representation, with the ability of being seen and interpreted by means of variety of diagrams, charts, tables, infographics and other forms of visual factors [133]. For example, the activities have characteristic of geolocation which can be projected onto a map, while the resources of knowledge can also be converted into the map. A map-based management and visual analysis method will largely benefit the users and the researchers from taking advantages of the Big data in education.

The authors in [134] proposed a novel map-based method to manage and analyze the mobile learning in Big education data. They retrieved the geographic location information from the GPS for the activities of participants recorded by the mobile learning systems and projected the data onto a map with a geographic reference and projection parameters. The layers of the new generated map can be subsequently integrated with an open map service like Google Map or Baidu Map. The learning activities and resources can be described as points, lines or polygons in the form of vector on the map. The map-based representations provide new methods (e.g. the map browsing) to perform exploration of learning practices. With their approach, the activities of users scattered among the space are reorganized on a geographic map with location changes in time series, and the resources are geotagged with the information from the developers or adopters, which are converted to a map style according to their hierarchical structures. The authors performed experiments using mobile learning data from the platform named M-starC of Central China Normal University (CCNU), which allows participants to use a mobile learning application for the access of the learning resources. Their experiment aimed to analyze the personal learning patterns. Classes were obtained from the data using the k-means method. The clusters revealed the spatial patterns of the individual who learned during the test duration. They also analyzed the group learning patterns from mobile learners and the location distribution.

The authors in [135] developed a novel approach, *Be the Data*, which exploits embodiment in visual analytics to invoke experiential learning. The authors designed and proposed a visual analytics approach to teach students about exploring alternative two-dimensional (2D) projections of high dimensional data points using weighted multi-dimensional scaling. In their approach, each student embodies a data point, and the position of students in a physical space represents a 2D projection of the high-dimensional data. Students physically move within the room with respect to each other to collaboratively construct alternative projections and receive visual feedback about relevant data dimensions. The approach exploits a large interactive room called the Cube and includes a large overhead display, a vision-based motion tracking system, and a software system for direct manipulation of high-dimensional data. To use the system, a group of students enter the Cube and embody virtual data points by wearing trackable hats which detect the locations of students in real-time. Their experimental findings indicate that *Be the Data* approach provided the engagement to enable students to quickly learn about high-dimensional data and analysis processes despite their minimal prior knowledge. They identified student data analytical strategies that employ this form of embodiment and found both qualitative and quantitative evidence of student improvement in understanding high-dimensional data. Visual Analytics approaches can also be usefully employed in MOOC. For example, VisMOOC [136] is an interactive visual analytics system, which can analyze video clickstream data by using a seeking diagram, PeakVizor [137] uses correlation view and flow view to uncover spatial and temporal information of peaks in video clickstreams from MOOC, and DropoutSeer system [138] uses timeline view by stack timelines and glyph to uncover the participants' learning activities and patterns, which can also predict the dropout.

F. IMMERSIVE LEARNING & ANALYTICS

The emergence of immersive learning approaches enabled by virtual reality (VR) technologies have given instructors and educators more flexibility and tools in designing active-based learning environments. Immersive learning techniques use computer graphics and human-computer interaction technologies to create simulated virtual worlds in which student learning can take place by employing suitable pedagogical approaches to create virtual worlds where learners could learn collaboratively [139], [140]. For example, the Second Life virtual world enables learners to create avatars in the virtual world for interaction with virtual objects and virtual environments [141]. Compared to traditional learning environments, immersive learning environments allow learners to explore problems and experience solutions in the virtual environment through experiential learning. The authors in [142] proposed an empirical study of designing and evaluating an immersive learning experience for a MOOC termed the VirtualHK MOOC. The authors work showed that immersive learning experience may not directly impact the knowledge gain of learning but can improve the overall learning experience in better motivating learners and making the learning more enjoyable. The student feedback and sentiment analysis showed that 52.73% of the learners gave "positive" comments and 47.27% gave "neutral" comments for the immersive learning experience.

G. SOCIAL MEDIA ANALYTICS

Student interactions and informal conversations on social media (e.g. Twitter, Facebook) give useful insights into their educational experiences, emotions and concerns about the learning process. However, data collection and analytics from social media data can be challenging due to the complexity. The collection of social media data has been presented in the previous section (Section IV). Normally, the student learning experiences acquired from social media content would require human interpretation. However, the growing scale of data volume and variety demands automatic data analytics techniques. This section focuses on a brief of mining social media data such as Twitter, followed by the inductive content analysis which frequently used in social media analytics and prominent themes. The previous section (Section IV) only presents the reviews of education data mining research. Here we give some examples of studies on Twitter from the fields of data mining, machine learning and natural language processing for education models and algorithms. The authors in [37] presented a work on mining social media data for understanding student learning experience from Twitter posts at Purdue University. The authors conducted a qualitative analysis taken from 25,000 tweets from engineering students and implemented a classification algorithm for tweets reflecting the student's problems. Their work presented a methodology that showed how data from social media can be used to provide insight into student learning experiences. The proliferation of multimedia technology in social learning spaces allows student emotions and sentiments to be captured and automatically classified from audio-visual devices such as web-cameras and microphones [149].

VII. CHALLENGES FOR BIG DATA IN EDUCATION AND LEARNING ANALYTICS

This section presents challenges for Big data in education and learning analytics from two perspectives: (1) social challenges; and (2) technological challenges. The technological and practical challenges are illustrated by giving an example for utilizing graph analytics for a university-based learning analytics scenario.

A. SOCIAL CHALLENGES

As in many fields where large amounts of data are being collected, there are also several important social challenges including privacy, ethical, security and safety issues to be addressed for Big education data. The authors in [144] considered a scenario where learning analytics (LA) could be used to track students and their performance could be flagged to deny a student access to future education programs based on the pre-conceived student ability for institution decision-making leading to unintended outcomes. The authors in [145] remarked that LA presents significant student privacy challenges for higher education institutions. In their work, the authors also posited four proponents that LA must justify in relation to the use of student data: (1) LA systems should provide controls for differential access to private student data; (2) Institutions must be able to justify their data collection using specific criteria; (3) The actual or perceived positive consequences of LA may not be equally beneficial for all students. A full accounting is required of how benefits are distributed between institutions and students and among students; and (4) Students should be made aware of collection and use of their data and permitted reasonable choices regarding collection and use of that data. The authors in [146] remarked that privacy and data protection are major stumbling blocks for a data-driven educational future. In this work, the authors proposed three principles to guide the practical deployment of LA and Big education data systems: (1) Privacy and data protection in LA are achieved by negotiating data sharing with each student; (2) How the educational institution will use data and act upon the insights of analysis should be clarified in close dialogue with the students; and (3) In negotiating privacy and data protection measures with students, schools and universities should use this opportunity to strengthen their personal data literacies.

B. TECHNOLOGICAL CHALLENGES

There are several technological opportunities and challenges for employing Big data in education and learning analytics due to the large and increasing amounts of online education data. As discussed in Section V, Big education systems would require access to a high-performance computational infrastructure which can handle a large amount of data for capture, storage, processing and visualization. There are also several issues and considerations for practical deployment of Big education data systems due to lack of interoperability of institutional data systems and different forms of data storage in disparate databases [143]. The absence of cross-institutional policies for data sharing and integration creates another major challenge to be addressed for Big education systems [38]. To illustrate some technological challenges and the usefulness and potential of exploiting cross-institutional Big education data, we performed an investigation for practical deployment of a Big education data system across some institutions in Australia. The objective of the system is to detect the unobserved prerequisite dependencies among online courses for different universities in Australia. This system would be useful for students to infer prerequisite relations among courses from different providers (e.g. universities, MOOC) to chart their learning pathways.

Our approach is based on graph-based analytics like the techniques proposed by [127], [128]. The graph-based

TABLE 3. Data Statistics for crawled university subject data.

Universities	Subjects	Prerequisites	Key words
ACU	1940	998	6602
ANU	2029	1576	10420
BU	651	139	5073
Total key words after merging	13108		

TABLE 4. Performance using MAP for cross-institution subject data.

Trained	Tested dataset from		
dataset from	ACU	ANU	BU
ACU	0.46	0.02	0.06
ANU	0.02	0.14	0.03
BU	0.01	0.01	0.39

TABLE 5. Performance using AUC for cross-institution subject data.

Trained	Tested dataset from		
dataset from	ACU	ANU	BU
ACU	0.96	0.56	0.65
ANU	0.52	0.67	0.48
BU	0.55	0.51	0.87

analytics approach was selected due to its effectiveness for cross-university transfer learning where courses may come from different providers and across institutions. For an initial investigation, we performed data collection from three universities in Australia (Australian National University – ANU, Australian Catholic University – ACU, and Bond University – BU) by regular web scraping techniques using Python on the respective university subject data available on the Internet. One challenge that was faced in the data collection process was to scrape the dynamic generated subject data from ANU, where we used Selenium to complete this task. Another challenge was to clean the raw data. We used standard text preprocessing methods to remove stop words (e.g. ''and'', ''is'', ''the'') and rare words with a training set frequency of 1. The raw data was cleaned by four methods including: (1) Conversion of data to lowercase; (2) Tokenization; (3) Word stemming; and (4) Removal of stop words and symbols. The Bag of Words (BoW) was used for modeling the cleaned data, with the extracted data statistics and total key words for the crawled university subject data as shown in Table 3. The BoW approach is a representation technique originating from Natural Language Processing (NLP) which is commonly used to extract features from text documents and other objects [150]. The ''Subjects'' field list the number of subjects in each university. The ''Prerequisites'' field shows the number of dependencies among subjects in the university. The ''Key Words'' field delivered the number of key words extracted from the ''Subject description'' in every university. And the ''Total key words'' field show the number of key words after merging of key words from the three universities. The generated links (Prerequisites) and the BoW model were imported into Matlab by libSVM as inputs into the graph-based algorithms. Two metrics (Mean Average Precision (MAP) and Area Under the Curve (AUC)) were used to evaluate the performance of the algorithms. The experiments were carried out on a workstation with an Intel i7-6800k CPU and 32GB RAM under Ubuntu 18.04 LTS. Table 4 and Table 5 shows the performance of the graph-based analytics among the three universities. The resulting subject data were split into training sets and test sets.

The models were trained using the dataset from one university, and then tested using the dataset from a different university. Some observations can be made from the results in Table 4 and Table 5. The performance of the AUC scores are higher than the MAP scores for the within-university performance (shown as the diagonals in the tables). On the one hand, the AUC gives equal weight to the predicted true positives. There may be different paths to achieve a goal and the AUC metric may evaluate them as giving similar performance. On the other hand, the MAP metric sorts true positives to higher positions of ranked lists to rank true positives higher than false positives to achieve a high MAP score. The score of MAP will be low if it fails to get higher rank on true positives. For the Table 4, the MAP may not always get higher rank on true positives. On cross-university performance, the scores of AUC were still higher than MAP, while lower than the AUC performance of within-university. The scores of MAP were lower than the within-university one. The performance of the cross-university was lower than the within-university performance due to the usage of different key words (labels) as inputs.

VIII. CONCLUSION

This paper has presented a comprehensive survey of research works on Big education data including the data sources, data collection, technological aspects, data analytics and challenges. The different sources for input into Big education data systems have also been discussed including learning management systems (LMS), open educational resources (OER), MOOC, social media and linked data. A classification of the various approaches for analytics have also been given which includes predictive analytics, learning analytics (collaborative/interactive learning, behaviour learning, personalized learning, and social learning), recommendation systems, graph analytics, visual analytics, social media analytics and immersive learning and analytics. The paper has also discussed social (privacy and ethical issues) and technological challenges for Big education data to be addressed for future research. Investigations for a cross-institution learning analytics scenario have also been given to illustrate the usefulness and technological challenges faced for practical deployment of Big education systems. The research area of Big education data is constantly evolving and amongst other sources, readers can refer to learning forums such as LAK (Learning Analytics & Knowledge Conference), Learning@Scale, AIED (Artificial Intelligence in Education) and periodicals such as JEDM (Journal of Educational Data Mining), IEEE Transactions on Learning Technologies for the latest research.

REFERENCES

[1] A. S. Alblawi and A. A. Alhamed, "Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics," in *Proc. IEEE Conf. Big Data Analytics (ICBDA)*, Nov. 2017, pp. 124–129.

[2] P. Michalik, J. Stofa, and I. Zolotova, "Concept definition for big data architecture in the education system," in *Proc. IEEE 12th Int. Symp. Appl. Mach. Intell. Informat. (SAMI)*, Jan. 2014, pp. 331–334.

[3] R. Machova, J. Komarkova, and M. Lnenicka, "Processing of big educational data in the cloud using apache Hadoop," in *Proc. Int. Conf. Inf. Soc. (i-Soc.)*, Oct. 2016, pp. 46–49.

[4] M.-S. Lee, E. Kim, C.-S. Nam, and D.-R. Shin, "Design of educational big data application using spark," in *Proc. 19th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2017, pp. 355–357.

[5] Q. Zheng, H. He, T. Ma, N. Xue, B. Li, and B. Dong, "Big log analysis for E-Learning ecosystem," in *Proc. IEEE 11th Int. Conf. e-Bus. Eng.*, Nov. 2014, pp. 258–263.

[6] D. Marjanovic, M. Milovanovic, and B. Radenkovic, "Hadoop infrastructure for education," in *Proc. 14th Int. Symp. New Bus. Models Sustain. Competitiveness*, 2014, pp. 365–370.

[7] C. Zhenyu, "The application of big data in higher vocational education based on holland vocational interest theory," in *Proc. Int. Conf. Ind. Informat. Comput. Technol., Intell. Technol., Ind. Inf. Integr. (ICIICII)*, Dec. 2017, pp. 37–40.

[8] H. Wang, Q. Wang, and W. Wang, "Text mining for educational literature on big data with Hadoop," in *Proc. IEEE Int. Conf. Smart Cloud (SmartCloud)*, Sep. 2018, pp. 166–170.

[9] R. Swathi, N. P. Kumar, L. Kirankranth, L. S. Madhav, and R. Seshadri, "Systematic approach on big data analytics in education systems," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2017, pp. 420–423.

[10] J. Chen, J. Tang, Q. Jiang, Y. Wang, C. Tao, X. Zhang, and J. Liao, "Research on architecture of education big data analysis system," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 601–605.

[11] M. W. Rodrigues, S. Isotani, and L. E. Zárate, "Educational data mining: A review of evaluation process in the e-learning," *Telematics Informat.*, vol. 35, no. 6, pp. 1701–1717, Sep. 2018.

[12] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, Jul. 2007.

[13] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.

[14] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010

[15] R. Sachin and M. Vijay, "A survey and future vision of data mining in educational field, advanced computing communication technologies (ACCT)," in *Proc. 2nd Int. Conf.*, Jan. 2012, pp. 96–100.

[16] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," in *Proc. 9th Int. Conf. Cognit. Sci.*, vol. 97, pp. 320–324, Nov. 2013.

[17] R. Jindal and M. D. Borah, "A survey on educational data mining and research trends," *Int. J. Database Manage. Syst.*, vol. 5, no. 3, pp. 53–73, Jun. 2013.

[18] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[19] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics Informat.*, vol. 37, pp. 13–49, Apr. 2019.

[20] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Hum. Behav.*, vol. 89, pp. 98–110, Dec. 2018.

[21] A. Peña-Ayala, "Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 3, May 2018, Art. no. e1243.

[22] R. Ferguson and D. Clow, "Where is the evidence?: A call to action for learning analytics," in *Proc. 7th Int. Learn. Analytics Knowl. Conf.*, Mar. 2017, pp. 56–65.

[23] P. Leitner, M. Khalil, and M. Ebner, "Learning analytics in higher education—A literature review," in *Learning Analytics: Fundaments, Applications, and Trends* (Studies in Systems, Decision and Control), Vol. 94, A. Peña-Ayala, Ed. Cham, Switzerland: Springer, 2017, pp. 1–23.

[24] Z. K. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Educ. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, Oct. 2014.

[25] L. K. Chew, "Using xAPI and learning analytics in education," in *Elearning Forum Asia*, 2016, pp. 13–15.

[26] O. Bohl, J. Scheuhase, R. Sengler, and U. Winand, "The sharable content object reference model (SCORM)—A critical review," in *Proc. Int. Conf. Comput. Edu.*, Dec. 2002, pp. 950–951.

[27] J. P. Leal and R. Queirós, "Using the learning tools interoperability framework for LMS integration in service oriented architectures," *Technol. Enhanced Learn. Tech-Educ.*, to be published.

[28] M. Dougiamas and P. Taylor, "Moodle: Using learning communities to create an open source course management system," in *Proc. EdMedia+ Innovate Learn., Assoc. Advancement Comput. Educ. (AACE)*, 2003, pp. 171–178.

[29] *The Top Open Source Learning Management Systems*. Accessed: Feb. 2020. [Online]. Available: https://elearningindustry.com/top-open-source-learning-management-systems

[30] M. H. Mohamed and M. Hammond, "MOOCs: A differentiation by pedagogy, content and assessment," *Int. J. Inf. Learn. Technol.*, vol. 35, no. 1, pp. 2–11, Jan. 2018.

[31] A. Agrawal, A. Kumar, and P. Agrawal, "Massive open online courses: EdX. org, Coursera. com and NPTEL, a comparative study based on usage statistics and features with special reference to India," INFLIBNET Centre, Tech. Rep., 2015.

[32] S. I. El Ahrache, H. Badir, Y. Tabaa, and A. Medouri, "Massive open online courses: A new dawn for higher education," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 5, p. 323, 2013.

[33] [Online]. Available: http://sociallearningcommunity.com/10-of-the-best-mooc-providers/

[34] [Online]. Available: https://en.unesco.org/events/experts-meeting-defining-open-educational-resources-oer-indicators

[35] [Online]. Available: http://discourse.col.org/t/what-are-examples-of-oer/27

[36] J. C. Taylor, "Open courseware futures: Creating a parallel universe," *e-JIST*, vol. 10, no. 1, pp. 1–7, 2007.

[37] X. Chen, M. Vorvoreanu, and K. P. C. Madhavan, "Mining social media data for understanding students' learning experiences," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 246–259, Jul. 2014.

[38] A. Dix, "Challenge and potential of fine grain, cross-institutional learning data," in *Proc. 3rd ACM Conf. Learn. Scale L@S*, 2016, pp. 261–264.

[39] C. K. Pereira, S. W. M. Siqueira, B. P. Nunes, and S. Dietze, "Linked data in education: A survey and a synthesis of actual research and future challenges," *IEEE Trans. Learn. Technol.*, vol. 11, no. 3, pp. 400–412, Jul. 2018.

[40] D. Taibi and S, Dietze, "Fostering analytics on learning analytics research: The LAK dataset," in *Proc. CEUR Workshop*. vol. 974, 2013, pp. 5–7.

[41] R. Meymandpour and J. G. Davis, "Ranking universities using linked open data," *J. Stud. Int. Educ.*, vol. 18, no. 2, pp. 318–327, 2007.

[42] B. E. Penteado, "Correlational analysis between school performance and municipal indicators in Brazil supported by linked open data," in *Proc. 25th Int. Conf. Companion World Wide Web WWW Companion*, 2016, pp. 507–512.

[43] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in *Proc. IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT)*, Nov. 2015, pp. 1–5.

[44] C. Keßler, M. d'Aquin, and S. Dietze, "Linked data for science and education," *Semantic Web*, vol. 4, no. 1, pp. 1–2, 2013.

[45] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a Web of open data," In *The Semantic Web*. Berlin, Germany: Springer, 2007, pp. 722–735.

[46] K. Bollacker, R. Cook, and P. Tufts, "Freebase: A shared database of structured general human knowledge," in *Proc. AAAI* vol. 7, Jul. 2007, pp. 1962–1963.

[47] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum, "YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, Oct. 2016, pp. 177–185.

[48] N. Bassiliades, "Collecting university rankings for comparison using Web extraction and entity linking techniques," in *Information and Communication Technologies in Education, Research, and Industrial Applications* (Communications in Computer and Information Science), vol. 469, 2014, pp. 23–46.

[49] J. Robinson, J. Stan, and M. Ribière, "Using linked data to reduce learning latency for e-book readers," in *Proc. Extended Semantic Web Conf.*, 2012, pp. 28–34.

[50] L. D. Rubenstein, "Using TED talks to inspire thoughtful practice," *Teacher Educator*, vol. 47, no. 4, pp. 261–267, Oct. 2012.

[51] [Online]. Available: http://data.linkededucation.org/linkedup/catalog/

[52] R. A. Huebner, "A survey of educational data-mining research," *Res. Higher Educ. J.*, vol. 19, no. 4, pp. 1–13, 2013.

[53] P. Guleria and M. Sood, "Data mining in education: A review on the knowledge discovery perspective," *Int. J. Data Mining Knowl. Manage. Process*, vol. 4, no. 5, pp. 47–60, Sep. 2014.

[54] S. Yu, D. Yang, and X. Feng, "A big data analysis method for online education," in *Proc. 10th Int. Conf. Intell. Comput. Technol. Autom. (ICICTA)*, Oct. 2017, pp. 291–294.

[55] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, Jan. 2015.

[56] S. A. Noghabi, K. Paramasivam, Y. Pan, N. Ramesh, J. Bringhurst, I Gupta, and R. H. Campbell, "Samza: Stateful scalable stream processing at LinkedIn," *Proc. VLDB Endowment*, vol. 10, no. 12, pp. 1634–1645, Aug. 2017.

[57] S. Roy and S. N. Singh, "Emerging trends in applications of big data in educational data mining and learning analytics," in *Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng. Confluence*, Jan. 2017, pp. 193–198.

[58] L. Cen, D. Ruta, and J. Ng, "Big education: Opportunities for big data analytics," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 502–506, doi: 10.1109/ICDSP.2015.7251923.

[59] M. S. Vyas and R. Gulwani, "Predictive analytics for e learning system," in *Proc. Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2017, pp. 1–4.

[60] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, pp. 36–51, Nov. 2018.

[61] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Model. User-Adapted Interact.*, vol. 19, no. 3, pp. 243–266, Aug. 2009.

[62] M. Jose, P. S. Kurian, and V. Biju, "Progression analysis of students in a higher education institution using big data open source predictive modeling tool," in *Proc. 3rd MEC Int. Conf. Big Data Smart City (ICBDSC)*, Mar. 2016, pp. 1–5.

[63] B. Kumar and S. Pal, "Mining educational data to analyze students performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 1–8, 2012.

[64] M. H. Abdous, H. Wu, and C. J. Yen, "Using data mining for predicting relationships between online question theme and final grade," *J. Educ. Technol. Soc.*, vol. 15, no. 3, p. 77, 2012.

[65] A. Wolff, Z. Zdrahal, D. Herrmannova, and P. Knoth, "Predicting student performance from combined data sources," in *Educational Data Mining*, 2013.

[66] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use moodle courses," *Comput. Appl. Eng. Edu.*, vol. 21, no. 1, pp. 135–146, Mar. 2013.

[67] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting student performance: A statistical and data mining approach," *Int. J. Comput. Appl.*, vol. 63, no. 8, pp. 35–39, 2012.

[68] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Comput. Hum. Behav.*, vol. 47, pp. 168–181, Jun. 2015.

[69] M. Nasiri, B. Minaei, and F. Vafaei, "Predicting GPA and academic dismissal in LMS using educational data mining: A case mining," in *Proc. 6th Nat. 3rd Int. Conf. E-Learn. E-Teach.*, Feb. 2012, pp. 53–58.

[70] K. Bunkar, U. K. Singh, B. Pandya, and R. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification," in *Proc. 9th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Sep. 2012, pp. 1–5.

[71] [Online]. Available: web.ysu.edu/gen/ysu_generated_bin/documents/basic_module/Key_Causes_of_Student_AttritionComprehensive_Retention_Plan.pdf

[72] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in edx MOOCs," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2016, pp. 440–443.

[73] R. Kanth, M.-J. Laakso, P. Nevalainen, and J. Heikkonen, "Future educational technology with big data and learning analytics," in *Proc. IEEE 27th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2018, pp. 906–910.

[74] A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," in *Proc. Int. Conf. Soft-Comput. Netw. Secur. (ICSNS)*, Feb. 2015, pp. 1–7.

[75] W. L. Cambruzzi, S. J. Rigo, and J. L. Barbosa, "Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach," *J. UCS*, vol. 21, no. 1, pp. 23–47, 2015.

[76] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, Feb. 2016.

[77] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting students drop out: A case study," *Educ. Data Mining*, to be published.

[78] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Int. J. Speech Technol.*, vol. 38, no. 3, pp. 315–330, Apr. 2013.

[79] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting students drop out: A case study," presented at the Educ. Data Mining, Jul. 2009.

[80] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelinsky, "Predicting drop-out from social Behaviour of students," *Int. Educ. Data Mining Soc.*, to be published.

[81] Z. Wang, C. Zhu, Z. Ying, Y. Zhang, B. Wang, X. Jin, and H. Yang, "Design and implementation of early warning system based on educational big data," in *Proc. 5th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2018, pp. 549–553.

[82] T. Denley, "Degree compass: A course recommendation system," *Educause Rev. Online*, pp. 1–5, Jun. 2013.

[83] P. Guleria and M. Sood, "Big data analytics: Predicting academic course preference using Hadoop inspired mapreduce," in *Proc. 4th Int. Conf. Image Inf. Process. (ICIIP)*, Dec. 2017, pp. 1–4.

[84] A. Pejic and P. S. Molcer, "Exploring data mining possibilities on computer based problem solving data," in *Proc. IEEE 14th Int. Symp. Intell. Syst. Informat. (SISY)*, Aug. 2016, pp. 171–176.

[85] L. Cen, D. Ruta, L. Powell, and J. Ng, "Learning alone or in a group - an empirical case study of the collaborative learning patterns and their impact on student grades," in *Proc. Int. Conf. Interact. Collaborative Learn. (ICL)*, Dec. 2014.

[86] Á. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Á. Conde-González, and Á. Hernández-García, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning," *Comput. Hum. Behav.*, vol. 31, pp. 542–550, Feb. 2014.

[87] A. van Leeuwen, J. Janssen, G. Erkens, and M. Brekelmans, "Supporting teachers in guiding collaborating students: Effects of learning analytics in CSCL," *Comput. Edu.*, vol. 79, pp. 28–39, Oct. 2014.

[88] J. Janssen, G. Erkens, and G. Kanselaar, "Visualization of agreement and discussion processes during computer-supported collaborative learning," *Comput. Hum. Behav.*, vol. 23, no. 3, pp. 1105–1125, May 2007.

[89] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education," *Comput. Edu.*, vol. 96, pp. 42–54, May 2016.

[90] Á. Fidalgo-Blanco, M. L. Sein-Echaluce, F. J. García-Peñalvo, and M. Á. Conde, "Using learning analytics to improve teamwork assessment," *Comput. Hum. Behav.*, vol. 47, pp. 149–156, Jun. 2015.

[91] P. Williams, "Assessing collaborative learning: Big data, analytics and university futures," *Assessment Eval. Higher Edu.*, vol. 42, no. 6, pp. 978–989, Aug. 2017.

[92] L. dos Santos Machado and K. Becker, "Distance education: A Web usage mining case study for the evaluation of learning sites," in *Proc. 3rd IEEE Int. Conf. Adv. Technol.*, Jul. 2003, pp. 360–361.

[93] L. Wang, J. Li, L. Ding, and P. Li, "E-learning evaluation system based on data mining," in *Proc. 2nd Inf. Eng. Electron. Commerce (IEEC)*, Jul. 2010, pp. 1–3.

[94] V. Pascual-Cid, L. Vigentini, and M. Quixal, "Visualising virtual learning environments: Case studies of the Website exploration tool," in *Proc. 14th Int. Conf. Inf. Visualisation*, Jul. 2010, pp. 149–155.

[95] I. L. M. Ricarte, G. R. F. Junior,"A methodology for mining data from computer-supported learning environments," *Informática na educação: Teoria Prática*, vol. 14, no. 2, pp. 83–94, 2011.

[96] J. S. Kinnebrew, J. R. Segedy, and G. Biswas, "Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments," *IEEE Trans. Learn. Technol.*, vol. 10, no. 2, pp. 140–153, Apr. 2017.

[97] A. Nussbaumer, E.-C. Hillemann, C. Gütl, and D. Albert, "A competence-based service for supporting self-regulated learning in virtual environments," *J. Learn. Anal.*, vol. 2, no. 1, pp. 101–133, 2015.

[98] J. L. Sabourin, B. W. Mott, and J. C. Lester, "Early prediction of student self-regulation strategies by combining multiple models," *Int. Educ. Data Mining Soc.*, to be published.

[99] K. Pietrosanti. *When E-Learning Technologies Embrace Big Data*. Accessed: Feb. 2020. [Online]. Available: https://www.docebo.com/2013/12/06/when-elearning-technologiesembrace-big-data-2/

[100] K. Habitzel, T. D. Mrk, B. Stehno, and S. Prock, "Microlearning: Emerging concepts, practices and technologies after e-learning," *Proc. Microlearning Learn. Work. New Media*, vol. 5, no. 3, 2006.

[101] R. Ferguson and S. B. Shum, "Social learning analytics: Five approaches," presented at the Proc. 2nd Int. Conf. Learn. Anal. Knowl., 2012.

[102] E. Duval, "Attention please!: Learning analytics for visualization and recommendation," presented at the Proc. 1st Int. Conf. Learn. Anal. Knowl., 2011.

[103] C.-M. Chen, C.-M. Hong, and C.-C. Chang, "Mining interactive social network for recommending appropriate learning partners in a Web-based cooperative learning environment," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, Sep. 2008, pp. 642–647.

[104] E. A. Heathcote and S. P. Dawson, "Data mining for evaluation, benchmarking and reflective practice in a LMS," presented at the E-Learn World Conf. E-Learn. Corporate, Government, Heathcare Higher Educ., Vancouver, BC, Canada, Oct. 2005.

[105] [Online]. Available: https://xapi.com/overview/

[106] M. Manso-Vazquez, M. Caeiro-Rodriguez, and M. Llamas-Nistal, "XAPI-SRL: Uses of an application profile for self-regulated learning based on the analysis of learning strategies," in *Proc. IEEE Frontiers Edu. Conf. (FIE)*, Oct. 2015, pp. 1–8.

[107] Y. Wu, S. Guo, and L. Zhu, "Design and implementation of data collection mechanism for 3D design course based on xAPI standard," *Interact. Learn. Environments*, pp. 1–18, Dec. 2019.

[108] A. Berg, M. Scheffel, H. Drachsler, S. Ternier, and M. Specht, "Dutch cooking with xAPI recipes: The good, the bad, and the consistent," in *Proc. IEEE 16th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2016, pp. 234–236.

[109] A. Nouira, L. Cheniti-Belcadhi, and R. Braham, "An enhanced xAPI data model supporting assessment analytics," *Procedia Comput. Sci.*, vol. 126, pp. 566–575, Jan. 2018.

[110] C. Ellis, "Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics," *Brit. J. Educ. Technol.*, vol. 44, no. 4, pp. 662–664, Jul. 2013.

[111] L. Cao, "Non-IID recommender systems: A review and framework of recommendation paradigm shifting," *Engineering*, vol. 2, no. 2, pp. 212–224, Jun. 2016.

[112] S. Dwivedi and V. S. K. Roshni, "Recommender system for big data in education," in *Proc. 5th Nat. Conf. E-Learn. E-Learn. Technol. (ELELTECH)*, Aug. 2017, pp. 1–4.

[113] Y. Hou, P. Zhou, J. Xu, and D. O. Wu, "Course recommendation of MOOC with big data support: A contextual online learning approach," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 106–111.

[114] M. Qbadou, I. Salhi, and K. Mansouri, "Towards an educational recommendation system based on big data techniques-case of Hadoop," in *Proc. 4th Int. Conf. Optim. Appl. (ICOA)*, Apr. 2018, pp. 1–5.

[115] L. Feng and G. Wei-wei, "Design and implementation of personalized recommendation system under big data platform," in *Proc. 11th Int. Conf. Intell. Comput. Technol. Autom. (ICICTA)*, Sep. 2018, pp. 291–294.

[116] M.-H. Hsu, "A personalized english learning recommender system for ESL students," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 683–688, Jan. 2008.

[117] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Comput. Sci.*, vol. 1, no. 2, pp. 2811–2819, 2010.

[118] O. C. Santos and J. G. Boticario, "Requirements for semantic educational recommender systems in formal E-learning scenarios," *Algorithms*, vol. 4, no. 2, pp. 131–154, 2011.

[119] O. R. Zaiane, "Building a recommender agent for e-learning systems," in *Proc. Int. Conf. Comput. Edu.*, Dec. 2002, pp. 55–59.

[120] J. Lu, "Personalized e-learning material recommender system," in *Proc. Int. Conf. Inf. Technol. Appl.*, 2004, pp. 374–379.

[121] F.-H. Wang and H.-M. Shao, "Effective personalized recommendation based on time-framed navigation clustering and association mining," *Expert Syst. Appl.*, vol. 27, no. 3, pp. 365–377, Oct. 2004.

[122] N. Baloian, P. Galdames, C. A. Collazos, and L. A. Guerrero, "A model for a collaborative recommender system for multimedia learning material," in *Proc. Int Conf. Collaboration Technol.*, Sep. 2004, pp. 281–288.

[123] C.-M. Chen, H.-M. Lee, and Y.-H. Chen, "Personalized e-learning system using item response theory," *Comput. Edu.*, vol. 44, no. 3, pp. 237–255, Apr. 2005.

[124] M. Gomez-Albarran and G. Jimenez-Diaz, "Recommendation and students' authoring in repositories of learning objects: A case-based reasoning approach," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 4, pp. 35–40, Oct. 2009.

[125] M. K. Khribi, M. Jemni, and O. Nasraoui, "Toward a hybrid recommender system for e-learning personalization based on Web usage mining techniques and information retrieval," in *Proc. World Conf. E-Learn. Corporate, Government, Healthcare Higher Educ.*, Oct. 2007, pp. 6136–6145.

[126] Y. Yang, H. Liu, J. Carbonell, and W. Ma, "Concept graph learning from educational data," in *Proc. 8th ACM Int. Conf. Web Search Data Mining WSDM*, 2015, pp. 159–168.

[127] H. Liu and Y. Yang, "Cross-graph learning of multi-relational associations," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2235–2243.

[128] W. Chen, C. G. Brinton, D. Cao, A. Mason-Singh, C. Lu, and M. Chiang, "Early detection prediction of learning outcomes in online short-courses via learning behaviors," *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 44–58, Jan. 2019.

[129] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Comput. Edu.*, vol. 122, pp. 119–135, Jul. 2018.

[130] J. Yoo, S. Yoo, C. Lance, and J. Hankins, "Student progress monitoring tool using treeview," presented at the ACM SIGCSE Bulletin, 2006.

[131] L. P. Macfadyen and P. Sorenson, "Using LiMS (the learner interaction monitoring system) to track online learner engagement and evaluate course design," presented at the Educ. Data Mining, Jun. 2010.

[132] J. Zeitz, N. Self, L. House, J. R. Evia, S. Leman, and C. North, "Bringing interactive visual analytics to the classroom for developing EDA skills," *J. Comput. Sci. Colleges*, vol. 33, no. 3, pp. 115–125, 2018.

[133] D. Zhou, H. Li, S. Liu, B. Song, and T. Hu, "A map-based visual analysis method for patterns discovery of mobile learning in education with big data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3482–3491.

[134] X. Chen, J. Zeitz Self, L. House, J. Wenskovitch, M. Sun, N. Wycoff, J. Robertson Evia, S. Leman, and C. North, "Be the data: Embodied visual analytics," *IEEE Trans. Learn. Technol.*, vol. 11, no. 1, pp. 81–95, Mar. 2018.

[135] C. Shi, S. Fu, Q. Chen, and H. Qu, "VisMOOC: Visualizing video clickstream data from massive open online courses," in *Proc. IEEE Pacific Visualizat. Symp. (PacificVis)*, Apr. 2015, pp. 159–166.

[136] Q. Chen, Y. Chen, D. Liu, C. Shi, Y. Wu, and H. Qu, "PeakVizor: Visual analytics of peaks in video clickstreams from massive open online courses," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 10, pp. 2315–2330, Oct. 2016.

[137] Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, and H. Qu, "DropoutSeer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol. (VAST)*, Oct. 2016, pp. 111–120.

[138] J. Herrington, T. C. Reeves, and R. Oliver, "Immersive learning technologies: Realism and online authentic learning," *J. Comput. Higher Edu.*, vol. 19, no. 1, pp. 80–99, Sep. 2007.

[139] Z. Pan, A. D. Cheok, H. Yang, J. Zhu, and J. Shi, "Virtual reality and mixed reality for virtual learning environments," *Comput. Graph.*, vol. 30, no. 1, pp. 20–28, Feb. 2006.

[140] S. C. Baker, R. K. Wentz, and M. M. Woods, "Using virtual worlds in education: Second Life as an educational tool," *Teach. Psychol.*, vol. 36, no. 1, pp. 59–64, Jan. 2009.

[141] H. H. S. Ip, C. Li, S. Leoni, Y. Chen, K.-F. Ma, C. H.-T. Wong, and Q. Li, "Design and evaluate immersive learning experience for massive open online courses (MOOCs)," *IEEE Trans. Learn. Technol.*, vol. 12, no. 4, pp. 503–515, Oct. 2019.

[142] B. Daniel, "Big data and analytics in higher education: Opportunities and challenges," *Brit. J. Educ. Technol.*, vol. 46, no. 5, pp. 904–920, Sep. 2015.

[143] B. K. Daniel, "Big data and data science: A critical review of issues for educational research," *Brit. J. Educ. Technol.*, vol. 50, no. 1, pp. 101–113, Jan. 2019.

[144] A. Rubel and K. M. L. Jones, "Student privacy in learning analytics: An information ethics perspective," *Inf. Soc.*, vol. 32, no. 2, pp. 143–159, Mar. 2016.

[145] T. Hoel and W. Chen, "Privacy and data protection in learning analytics should be motivated by an educational maxim—Towards a proposal," *Res. Pract. Technol. Enhanced Learn.*, vol. 13, no. 1, pp. 1–14, Dec. 2018.

[146] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014.

[147] R. H. L. Ip, L.-M. Ang, K. P. Seng, J. C. Broster, and J. E. Pratley, "Big data and machine learning for crop protection," *Comput. Electron. Agricult.*, vol. 151, pp. 376–383, Aug. 2018.

[148] K. P. Seng, L. M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 3–13, Jan./Mar. 2018.

[149] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[150] [Online]. Available: https://www.instructure.com/canvas/

[151] B. Flanagan and H. Ogata, "Learning analytics platform in higher education in Japan," *Knowl. Manage. E-Learn. (KM&EL)*, vol. 10, no. 4, pp. 469–484, Nov. 2018.

[152] M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, "Analysis of student behavior in learning management systems through a big data framework," *Future Gener. Comput. Syst.*, vol. 90, pp. 262–272, Jan. 2019.

[153] O. K. Akputu, K. P. Seng, Y. Lee, and L.-M. Ang, "Emotion recognition using multiple kernel learning toward E-learning applications," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–20, Jan. 2018.

KENNETH LI-MINN ANG (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from Edith Cowan University, Australia. He was an Associate Professor of networked and computer systems with the School of Information and Communication Technology (ICT), Griffith University. He is currently a Professor with the School of Science and Engineering, University of Sunshine Coast. His research interests include big data analytics, multimedia Internet-of-Things, embedded systems, wireless multimedia sensor systems, reconfigurable computing and the development of real-world computer systems, and machine learning. He has published over 180 articles in journals and international refereed conferences. He is a Fellow of the Higher Education Academy, U.K.

FENG LU GE received the B.Sc. degree in information engineering from the Dalian University of Technology, China, the M.Sc. degree from the University of Wollongong, Australia, and the Ph.D. degree from Charles Sturt University. He is currently an Engineer with Pacific Telecom & Navigation Ltd., Hong Kong. He was previously a Postdoctoral Researcher with Charles Sturt University. His research interests include data analytics, computer vision, and robotics.

KAH PHOOI SENG (Member, IEEE) received the B.Eng. and Ph.D. degrees from the University of Tasmania, Australia. She is currently an Adjunct Professor with the School of Engineering and Information Technology, UNSW. Before returning to Australia, she was a Professor and the Department Head of computer science and networked system with Sunway University. Before joining Sunway University, she was an Associate Professor with the School of Electrical and Electronic Engineering, Nottingham University. She has published over 230 articles in journals and international refereed conferences. She is the lead author of the book *Multimodal Analytics for Next-Generation Big Data Technologies and Applications*. Her research interests include data analytics, big data, machine learning, artificial intelligence (AI) and intelligent systems, the Internet of Things (IoT), multimodal signal processing, pervasive computing and sensor networks, HCI and affective computing, and mobile software development.