A survey of educational data-mining research

Richard A. Huebner
Norwich University

ABSTRACT

Educational data mining (EDM) is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. The discipline focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness. A literature review on educational data mining follows, which covers topics such as student retention and attrition, personal recommender systems within education, and how data mining can be used to analyze course management system data. Gaps in the current literature and opportunities for further research are presented.

Keywords: educational data mining, academic analytics, learning analytics, institutional effectiveness

INTRODUCTION

There is pressure in higher educational institutions to provide up-to-date information on institutional effectiveness (C. Romero & Ventura, 2010). Institutions are also increasingly held accountable for student success (Campbell & Oblinger, 2007). One response to this pressure is finding new ways to apply analytical and data mining methods to educationally related data. Even though data mining (DM) has been applied in numerous industries and sectors, the application of DM to educational contexts is limited (Ranjan & Malik, 2007). Researchers have found that they can apply data mining to rich educational data sets that come from course management systems such as Angel, Blackboard, WebCT, and Moodle. The emerging field of educational data mining (EDM) examines the unique ways of applying data mining methods to solve educationally related problems.

The recent literature related to educational data mining (EDM) is presented. Educational data mining is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data (Baker & Yacef, 2009). Researchers within EDM focus on topics ranging from using data mining to improve institutional effectiveness to applying data mining in improving student learning processes. There is a wide range of topics within educational data mining, so this paper will focus exclusively on ways that data mining is used to improve student success and processes directly related to student learning. For example, student success and retention, personalized recommender systems, and evaluation of student learning within course management systems (CMS) are all topics within the broad field of educational data mining.

Researchers interested in educational data mining established the *Journal of Educational Data Mining* (2009) and a yearly international conference that began in 2008. The EDM literature draws from several reference disciplines including data mining, learning theory, data visualization, machine learning and psychometrics (Baker & Yacef, 2009). Some of the earliest works are published in the *Conference on Artificial Intelligence in Education*, and the *International Journal of Artificial Intelligence in Education*. Interestingly, artificial intelligence is a large part of data mining, which is why we see early educational data mining papers in artificial intelligence related publications.

The purpose of this paper is to provide a survey of educational data mining research. Specific applications of educational data mining are delineated, which include student retention and attrition, personal recommender systems, and other data mining studies within course management systems. The paper concludes with identifying gaps in the current literature and recommendations for further research.

BACKGROUND OF DATA MINING

Big data is a term that describes the growth of the amount of data that is available to an organization and the potential to discover new insights when analyzing the data. IBM suggests big data spans three different dimensions, which include volume, velocity, and variety (IBM, 2012). Organizations have a challenge of sifting through all of that information, and need solutions to do so. Data mining can assist organizations with uncovering useful information in order to guide decision-making (Kiron, Shockley, Kruschwitz, Finch, & Haydock, 2012). Data mining is a series of tools and techniques for uncovering hidden patterns and relationships among data (Dunham, 2003). Data mining is also one step in an overall knowledge discovery process, where organizations want to discover new information from the data in order to aid in decision-making processes. Knowledge discovery and data mining can be thought of as tools for decision-making and organizational effectiveness. The complexity of data mining has led the data analytics community to establish a standard process for data mining activities.

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a life cycle process for developing and analyzing data mining models (Leventhal, 2010). The CRISP-DM process is important because it gives specific tips and techniques on how to move from understanding the business data through deployment of a data mining model. CRISP-DM has six phases, which include business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Leventhal, 2010). The benefits of CRISM-DM are that it is non-proprietary and software vendor neutral, and provides a solid framework for guidance in data mining (Leventhal, 2010). The model also includes templates to aid in analysis. This process is used in a number of educational data mining studies (Luan, 2002; Vialardi et al., 2011; Y.-h. Wang & Liao, 2011), but may not be explicitly stated as such.

Data mining has its roots in machine learning, artificial intelligence, computer science, and statistics (Dunham, 2003). There are a variety of different data mining techniques and approaches, such as clustering, classification, and association rule mining. Each of these approaches can be used to quantitatively analyze large data sets to find hidden meaning and patterns. Data mining is an exploratory process, but can be used for confirmatory investigations (Berson, Smith, & Thearling, 2011). It is different from other searching and analysis techniques in that data mining is highly exploratory, where other analyses are typically problem-driven and confirmatory.

While data mining has been applied in a variety of industries, government, military, retail, and banking, data mining has not received much attention in educational contexts (Ranjan & Malik, 2007). Educational data mining is a field of study that analyzes and applies data mining to solve educationally-related problems. Applying data mining this way can help researchers and practitioners discover new ways to uncover patterns and trends within large amounts of educational data.

BACKGROUND OF EDUCATIONAL DATA MINING

There are different ways that educational data mining is defined. Campbell and Oblinger (2007) defined academic analytics as the use of statistical techniques and data mining in ways that will help faculty and advisors become more proactive in identifying at-risk students and responding accordingly. In this way, the results of data mining can be used to improve student retention. Academic analytics focuses on processes that occur at the department, unit, or college and university level. This type of analysis does not focus on the details of each individual course, so it can be said that academic analytics has a macro perspective. Academic analytics can be considered a sub-field of educational data mining.

Baker and Yacef (2009) defined EDM as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (Baker & Yacef, 2009, p. 1). Their definition does not mention data mining, leaving researchers open to exploring and developing other analytical methods that can be applied to educationally related data. Also, many educators would not know how to use data mining tools, thus there is a need to make it easy for educators to conduct advanced analytics against data that pertains to them (such as online CMS data, etc.). One of the advantages to their research is that it provides a broad representation of the EDM field so far by discussing the prominent papers in the field. However, their research used the number of article citations as a way to evaluate growth of EDM. Perhaps future research can used a broader perspective when evaluating this discipline's growth.

In evaluating the above two definitions, educational data mining is a broader term that focuses on nearly any type of data in educational institutions, while academic analytics is specific to data related to institutional effectiveness and student retention issues. As noted earlier, the discipline relies on several reference disciplines and in the future, there will be additional growth in the interdisciplinary nature of EDM. As the discipline grows, researchers will need to refine the scope and definitions of EDM. At this early stage, it would be helpful to have a more thorough taxonomy of the different areas of study within EDM, even though a basic taxonomy has already been established by researchers (Baker & Yacef, 2009). One drawback to Baker and Yacef's taxonomy (2009) is that it does not address aspects of the clustering data mining task. Perhaps future research could expand on the clustering aspects of EDM.

The scope of educational data mining includes areas that directly impact students. For example, mining course content and the development of recommender systems (to be discussed later in this paper). Other areas within EDM include analysis of educational processes including admissions, alumni relations, and course selections. Furthermore, applications of specific data mining techniques such as web mining, classification, association rule mining, and multivariate statistics are also key techniques applied to educationally related data (Calders & Pechenizkiy, 2012). These data mining methods are largely exploratory techniques that can be used for prediction and forecasting of learning and institutional improvement needs. Also, the techniques can be used for modeling individual differences in students and provide a way to respond to those differences thus improve student learning (Corbett, 2001). Although, one question is how do institutions adopt educational data mining to improve institutional effectiveness?

In order for educational data mining to be successful, it is critical to have a solid data warehousing strategy. Guan et al. (2002) discussed how important it is to have meaningful information available for decision-makers within higher educational institutions. It is a challenge to get the information that decision makers need quickly and efficiently. Some of the primary drivers of initiating data warehouse projects include increased competitive landscape, and increased responsibilities of reporting to external stakeholders such as parents, board members, legislators and community leaders (Guan, Nunez, & Welsh, 2002).

Educational data mining can draw upon ideas from organizational data mining. Organizational data mining (ODM) focuses on assisting organizations with sustaining competitive advantage (Nemati & Barko, 2004). The key difference between DM and ODM is that ODM relies on organizational theory as a reference discipline (Nemati & Barko, 2004). Organizations that transform their data into useful information and knowledge, and do so efficiently, should gain tremendous benefits such as enhanced decision-making, increased competitiveness, and potential financial gains (Nemati & Barko, 2004). Therefore, the EDM field draws upon organizational theory as well. This is an important relationship because the focus of research within EDM can examine phenomena at different levels of analysis, from societal, organizational, unit, or individual level.

The type of research done within EDM focuses primarily on quantitative analyses, which is necessary because data mining employs statistics, machine learning, and artificial intelligence techniques. Many of the studies presented in this literature review are case studies where data mining projects were done at a specific institution, with a single institution's data. Qualitative techniques such as interviews and document analysis are also used to support case studies in EDM. The dominant research paradigm is quantitative, with results coming in the form of predictions, clusters or classifications, or associations. The drawback with some of the existing case studies is that the results are not necessarily generalizable to other institutions. This means that the results are highly associated with a specific institution at a specific time. Research in EDM should examine ways for data mining results to be more generalizable.

APPLICATIONS OF DATA MINING

A review of related literature in educational data mining follows. It focuses on how data mining is used for improving student success and processes directly related to student learning. Educational data mining research examines different ways that course management systems (CMS) data can be mined to provide new patterns of student behavior. Results can assist faculty and staff with improving learning and supporting educational processes, which in turn improve institutional effectiveness.

Student Retention and Attrition

Research has shown that data mining can be used to discover at-risk students and help institutions become much more proactive in identifying and responding to those students (Luan, 2002). Luan (2002) applied data mining as a way to predict what types of students would drop out of school, and then return to school later on. He applied classification and regression trees (C&RT) – a specific data mining technique – to educational data in order to predict which students are unlikely to return to school. In this case study, Luan applied both quantitative and qualitative research techniques to uncover student success factors. This research is important because it demonstrated the successful application of data mining tools to assist in student retention efforts. As noted earlier, the case study method for EDM may often produce results that are not generalizable. However, the process by which researchers apply the data mining can be generalized and used in other contexts. It is simply the results of the data mining models that may not be generalized.

In a related study, Lin (2012) applied data mining as a way to improve student retention efforts. Lin (2012) was able to generate predictive models based on incoming students' data. The models were able to provide short-term accuracy for predicting which types of students would benefit from student retention programs on campus. The research study found that certain machine learning algorithms can provide useful predictions of student retention (Lin, 2012).

Researchers at Bowie State University developed a system based on data mining that supports and improves retention (Chacon, Spicer, & Valbuena, 2012). Their system helps the institution identify and respond to at-risk students. Their research contributes meaningfully to the EDM literature because it demonstrates a successful implementation and use of data mining. Their work is highly representative of the discipline in that it follows a strict data mining process and is quantitative. Chacon et al.'s (2012) research supports other work done in applying data mining to student retention issues, such as Lin (2012) and Luan (2012), all with successful results. The work by Chacon et al. goes one step further than Lin and Luan, because the researchers were able to develop and implement their solution in a production environment. Bowie State University uses the system to aid in student retention efforts.

Data mining was used to assess the efficacy of a writing center in an effort to analyze student achievement and student progress to the next grade (Yeats, Reddy, Wheeler, Senior, & Murray, 2010). Their work demonstrated the ability to assess a specific educational support process, i.e., the writing center, in an effort to improve institutional effectiveness. Their research approach used a combination of quantitative work and case study analysis. The mixed-methods approach to data mining was helpful in understanding much more about the ways data mining can be used in an actual implementation. Their research results were not surprising in that it found students who attend writing centers tend to do better in their classes. The research by Yeats et al. (2010) took a different approach to analyzing student achievement in that it made the connection between writing center attendance and student grades. It did not make the link to student retention issues, but a future study could examine the relationship between these three concepts: writing center attendance, student grades, and retention.

In another study, three different data mining techniques were used to determine predictors of student retention. Yu, DiGangi, Jannesch-Pennell and Kaprolet (2010) applied classification trees, multivariate adaptive regression splines (MARS), and neural networks to educational data which resulted in finding transferred hours, residency, and ethnicity as critical elements in retention efforts (Yu, DiGangi, Jannasch-Pennell, & Kaprolet, 2010). Through this research, they also discovered that east coast students tend to stay enrolled longer than their west coast counterparts do.

Academic performance and student success can be predicted by using data mining techniques. One research team used data mining to classify students into three groups as early as they could in the academic year (Vandamme, Meskens, & Superby, 2007). The three groups included low-risk, medium risk, and high-risk students. The authors used several data mining techniques including neural networks, random forests, and decision trees. The student in the high risk group had a high probability of failing or dropping out of school. These types of studies are important in that they give faculty and staff a way to identify the at-risk students in a proactive way, because “once a student decides to leave, it is hard to convince them to stay” (discussion with Director of Institutional Effectiveness at Norwich University).

In a related study, researchers examined whether the demographic background of students had any influence on their performance (Yorke et al., 2005). Results from the study appeared inconclusive, potentially because of the type of analysis they did. Interestingly, the field of educational data mining is concerned with analytical methods, and not necessarily just data mining methods. Yorke et al. (2005) used Microsoft Excel for their analysis and mining data. The problem with this approach is that they discuss mining the data without really applying data mining techniques. It is clear that researchers should exercise more caution when using the phrase data mining, especially when they are not referring to data mining techniques. The drawback with the research Yorke et al. (2005) used these phrases, but never applied any classification, regression, or other data mining technique. This particular research demonstrates that researchers can still conduct data analyses by using Excel, but researchers should not mislead the reader when describing their approach. Contrary to the Yorke et al. (2005) study, a different research team noted that demographic characteristics are not significant predictors of student satisfaction or success (Thomas & Galambos, 2004). The results seem to report different findings related to student satisfaction or the prediction of student success. One can conclude there are significantly more factors that influence students’ success than what has been studied thus far.

Personal Learning Environments and Recommender Systems

Personal learning environments (PLEs) and personal recommendation systems (PRS) also directly relate to educational data mining. Personalized learning environments focus on providing the various tools, services, and artifacts so that the system can adapt to students' learning needs on the fly (Mödritscher, 2010). Much of the work done related to recommender systems is quantitative and is widely used in eCommerce. For example, Amazon.com uses recommender systems in order to customize the browsing experience for each user. Recommendations display related products that a consumer might purchase. Netflix also employs recommender systems to help its subscribers find the types of movies that they will probably like.

Recommender systems must be adapted when they are used in educational contexts because the recommendations should coincide with educational objectives. The reason is that it is not possible to apply existing recommender systems directly to educational data because they are highly domain dependent (Santos & Boticario, 2010). There are two significant challenges with respect to applying recommender systems in an educational context. First, the system must attempt to understand or determine the needs of learners. Second, there should be some way for faculty members to control recommendations for their learners (Santos & Boticario, 2010). Existing recommender systems in the educational domain typically do not address these concerns, which open up additional research opportunities for the EDM research community.

How can researchers and educational administrators use data mining to predict student performance? One research team examined this issue by applying recommender systems in an effort to improve student prediction results (Thai-Nghe, Drumond, Krohn-Grimberghe, & Schmidt-Thieme, 2010). This particular research study is one of the more quantitatively rigorous articles, probably more appropriate for computer science study, because it focuses on underlying algorithms and methods to improve recommender systems. However, the value of this study is that it provides an analysis of which analytical methods are more accurate when predicting student performance.

Recommendations for further learning exercises were made based on a student's web browsing behavior and improved student achievement. A data mining model was established that annotated browsing events with contextual factors, to produce new individualized content recommendations specifically for course management systems (F.-H. Wang, 2008). The results showed that data mining can deliver highly personalized content, based on browsing history and history of student achievement. This also improved student learning because students could move through the material at their own pace. The researchers also discovered that the contextual browsing model is much more effective than using association rule mining models.

Data mining was used in one study as a way to analyze users' preferences in interactive multimedia learning systems. The data mining clustering technique was used to place students into four main groups based on their preferences and computer experience (Chrysostomou, Chen, & Liu, 2009). Although the researchers used student preferences as a variable and determined that computer experience as a factor that influences preferences, it is unknown what other types of factors might influence preferences in an online learning environment. Future research could examine additional factors or demographics that contribute to student preferences, such as age, gender, or ethnicity.

Data mining was used in another study to provide learners with many recommendations to help them learn more effectively and efficiently. A methodology called frequent itemset mining was used to mine learner behavior patterns in an online course and subsequently, provide learners with different levels of recommendations rather than single ones that are produced from other recommender systems (Huang, Chen, & Cheng, 2007). This system assisted learners by providing them with highly individualized recommendations for improved learning efficiency.

A newer stream of research focuses on mobile learning environments. A study by Su, Tseng, Lin, and Chen (2011) applied data mining to help provide fast, dynamic, personalized learning content to mobile users. Mobile devices have very different requirements for managing content than standard PCs and web browsers (Su, Tseng, Lin, & Chen, 2011). They use data such as network conditions, hardware capabilities, and the user's preferences from their device. While this particular study is extremely technical, it demonstrates how mobile learning environments can benefit from data mining.

EDM AND COURSE MANAGEMENT SYSTEMS

A large number of researchers within EDM focus directly on course management systems and how they can be improved to support student learning outcomes and student success. One research team developed a simplified data mining toolkit that operates within the course management system and allows non-expert users to get data mining information for their courses (García, Romero, Ventura, & de Castro, 2011). In addition, a toolkit allows teachers to collaborate with each other and share results. This research is important because most data mining tools are complicated and require deep expertise in data mining tools, methods and processes, statistics, and machine learning algorithms. This study follows a typical data mining process, thus it is quantitative. The data mining process usually follows a pre-processing phase, then an application of specific data mining techniques, and then a post-processing phase. The research and application contributions will allow non-technical faculty to engage in educational data mining activities. It is clear that additional is needed in this area to make educational data mining tools more accessible to non-technical users.

Course management systems such as open source Moodle can be mined for usage data to find interesting patterns and trends in student online behavior. A systematic method for applying data mining techniques to Moodle usage data was established (Cristóbal Romero, Ventura, & García, 2008). The benefit to mining usage data is that it contains data about every user activity, such as testing, quizzes, reading, and discussion posts. Romero et al. (2008) discuss the importance of pre-processing the data and then discuss specifics on how to apply data mining techniques to Moodle data. Their research results demonstrated how straightforward it is to mine data, even if a reader does not have much experience in this area. The authors also use both Keel and Weka as their data mining software packages. These software programs are open source and are built on the Java language, so they are extendable as well.

Data mining can be used in such a way as to customize learning activities for each individual student. Data mining was used to adapt learning exercises based on students' progress through a course on English language instruction (Y.-h. Wang & Liao, 2011). Instead of having static course content, the course adapts to student learning, taking him or her through the course at his or her own pace. This was an effort to create significant and optimal learning experiences for each student, and was a success. This research could be applied to other types of courses where students begin a course with varying levels of competency, e.g., a computer programming course.

Data mining was used to assess complex student behaviors with respect to a three-week programming assignment. Blikstein (2011) found results that showed different types of student programming behaviors in an online course. These log files contained different types of events as each student completed them. The events included coding and non-coding activities in the online course. This quantitative data mining research helped discover different programming strategies used by students, and developed three programming behavior profiles: copy-and-pasters, mixed-mode, and self-sufficients (Blikstein, 2011).

In many online courses, discussion board posts are an important part of the learning experience. One research team used data mining as a strategy for assessing asynchronous discussion forums because it was challenging to manually assess the quality of the postings by each student (Dringus & Ellis, 2005). Their research attempts to answer the question of what kind of information is embedded in online discussion groups. The data mining results were used to assess student progress in an online course. One drawback with this approach is that non-technical faculty would not know how to apply data mining to get results for their students, thus there is a need to create tools that are accessible to non-technical faculty members.

Like Blikstein (2011), Dringus and Ellis (2005) analyze student behavior by applying data mining techniques. While the former examines programming activity behavior, the latter examines discussion board behavior. The analysis is different based upon the type of task or activity. For example, the DM analysis programming tasks in a course management system is going to be different than the DM analysis for discussion boards. Each data mining task is usually very specific and is used with a specific data set. However, may be more important to find ways of applying data mining to examine students' behavior in a broader sense, rather than analyzing a single aspect of their behavior within the CMS.

In an online educational environment, learner engagement is an important aspect of student success. Students' engagement with the course content can be analyzed using data mining techniques to determine if there are disengaged learners (Cocea & Weibelzahl, 2009). There were several factors that were revealed that contribute to predicting student disengagement, which included the speed at which students read through the pages, and the length of time spent on pages. Additionally, their study also determined that when students first logon to an online course, their behavior is quite erratic, probably because the student is learning how to use the course environment itself. Therefore, an analysis should take into account this type of behavior when producing data mining models.

One potential drawback to the use of online course management systems is that students can manipulate the system and avoid learning. Gaming is the idea that students attempt to circumvent properties of the system in order to make progress, while avoiding learning (Muldner, Burleson, Van de Sande, & Vanlehn, 2011). Some researchers are investigating what can be done to minimize gaming, and to make sure that students continue learning. Muldner et al. (2011) used data mining techniques including Bayesian methods (Naïve Bayes) and found that students, rather than the assignment or problem, was a better predictor of gaming. They also provided numerous recommendations for discouraging gaming. These include supplying extra or supplemental exercises, or the use of an intelligent agent that displays disapproval if gaming is detected within the system.

CONCLUSION AND FUTURE WORK

Educational data mining (EDM) is an area full of exciting opportunities for researchers and practitioners. This field assists higher educational institutions with efficient and effective ways to improve institutional effectiveness and student learning. Data mining is a significant tool for helping organizations enhance decision making and analyzing new patterns and relationships among a large amount of data. A broad sense of the types of research currently being conducted in EDM was presented, from applying data mining for understanding student retention and attrition to finding new ways of making personalized learning recommendations to each individual student. Many opportunities exist to study EDM from an organizational unit of analysis to individual course-levels of analysis. Some work is strategic in nature and some of the research is extremely technical. Overall, EDM draws upon several reference disciplines and continues to grow with the introduction of the Journal of Educational Data Mining and its related annual conference. These were established only in 2008, which indicates that the discipline is still in its infancy. It will be exciting to see how EDM develops over the coming years.

Bienkowski, Feng, and Means (2012) presented a thorough report on how educational data mining and learning analytics can enhance teaching and learning. The authors outlined compelling avenues for further research. These included:

- a focus on usability and impact of presenting learning data to instructors;
- development of decision support systems and recommendation systems that minimize instructor intervention;
- development of tools for protecting individual privacy while still advancing educational data mining; and
- development of models that can be used in multiple contexts.

Researchers have not addressed how data mining can be applied to plagiarism detection. Plagiarism is a topic that faculty become quite concerned with. Thus, it behooves us to develop predictive capability in plagiarism-related issues.

Future research can examine how widespread the adoption of educational data mining might be. Currently, it appears that research in this area is isolated and we do not know the exact extent of how institutions might be using data mining for enhancing student learning or improving related educational processes. Furthermore, we do not know if there are intentions to adopt EDM or any initiatives where institutions are considering adopting an EDM strategy. It would be interesting to determine if there are barriers that prevent institutions from establishing EDM initiatives. There are a few case studies on how EDM is applied to admissions and enrollment, but further work needs to be done because those case studies seem isolated from the mainstream EDM work.

REFERENCES

Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1*(1).

Berson, A., Smith, S., & Thearling, K. (2011). An Overview of Data Mining Techniques Retrieved November 28, 2011, from http://www.thearling.com/text/dmtechniques/dmtechniques.htm

Blikstein, P. (2011). *Using learning analytics to assess students' behavior in open-ended programming tasks*. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.
Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *SIGKDD Explor. Newsl., 13*(2), 3-6. doi: 10.1145/2207243.2207245
Campbell, J., & Oblinger, D. (2007). Academic analytics. Washington, DC: Educause.
Chacon, F., Spicer, D., & Valbuena, A. (2012). Analytics in Support of Student Retention and Success (Research Bulletin 3, 2012 ed.). Louisville, CO: Educause Center for Applied Research.
Chrysostomou, K., Chen, S. Y., & Liu, X. (2009). Investigation of Users' Preferences in Interactive Multimedia Learning Systems: A Data Mining Approach. *Interactive Learning Environments, 17*(2), 151-163.
Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Modeling and User - Adapted Interaction, 19*(4), 341-385. doi: 10.1007/s11257-009-9065-5
Corbett, A. T. (2001). *Cognitive Computer Tutors: Solving the Two-Sigma Problem*. Paper presented at the Proceedings of the 8th International Conference on User Modeling 2001.
Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*, 141-160.
Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Pearson Education.
García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education, 14*(2), 77-88. doi: 10.1016/j.iheduc.2010.07.006
Guan, J., Nunez, W., & Welsh, J. (2002). Institutional strategy and information support: the role of data warehousing in higher education. *Campus-Wide Information Systems, 19*(5), 168-174.
Huang, Y.-M., Chen, J.-N., & Cheng, S.-C. (2007). A Method of Cross-Level Frequent Pattern Mining for Web-Based Instruction. *Educational Technology & Society, 10*(3), 305-319.
IBM. (2012). What is big data? Retrieved May 16th, 2012, from http://www-01.ibm.com/software/data/bigdata/
Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2012). Analytics: The Widening Divide. *MIT Sloan Management Review, 53*(2), 1-22.
Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice, 12*(2), 137-153. doi: 10.1057/dddmp.2010.35
Lin, S.-H. (2012). Data mining for student retention management. *J. Comput. Sci. Coll., 27*(4), 92-99.
Luan, J. (2002). *Data Mining and Knowledge Management in Higher Education - Potential Applications*. Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada. http://eric.ed.gov/ERICWebPortal/detail?accno=ED474143
Mödritscher, F. (2010). Towards a recommender strategy for personal learning environments. *Procedia Computer Science, 1*(2), 2775-2782. doi: 10.1016/j.procs.2010.08.002

Muldner, K., Burleson, W., Van de Sande, B., & Vanlehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User Modeling and User - Adapted Interaction, 21*(1-2), 99-135. doi: 10.1007/s11257-010-9086-0

Nemati, H., & Barko, C. (2004). Organizational Data Mining (ODM): An Introduction. In H. Nemati & C. Barko (Eds.), *Organizational Data Mining* (pp. 1-8). London: Idea Group Publishing.

Ranjan, J., & Malik, K. (2007). Effective educational process: a data-mining approach. *VINE: The Journal of Information and Knowledge Management Systems, 37*(4), 502-515.

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40*(6), 601-618. doi: 10.1109/tsmcc.2010.2053532

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384. doi: 10.1016/j.compedu.2007.05.016

Santos, O. C., & Boticario, J. G. (2010). Modeling recommendations for the educational domain. *Procedia Computer Science, 1*(2), 2793-2800. doi: 10.1016/j.procs.2010.08.004

Su, J.-m., Tseng, S.-s., Lin, H.-y., & Chen, C.-h. (2011). A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments. *User Modeling and User - Adapted Interaction, 21*(1-2), 5-49. doi: 10.1007/s11257-010-9094-0

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science, 1*(2), 2811-2819. doi: 10.1016/j.procs.2010.08.006

Thomas, E. H., & Galambos, N. (2004). What Satisfies Students?: Mining Student-Opinion Data with Regression and Decision Tree Analysis. *Research in Higher Education, 45*(3), 251-269.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics, 15*(4), 405-419.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, Á. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User - Adapted Interaction, 21*(1-2), 217-248. doi: 10.1007/s11257-011-9098-4

Wang, F.-H. (2008). Content Recommendation Based on Education-Contextualized Browsing Events for Web-Based Personalized Learning. *Educational Technology & Society, 11*(4), 94-112.

Wang, Y.-h., & Liao, H.-C. (2011). Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications, 38*(6), 6480-6485. doi: 10.1016/j.eswa.2010.11.098

Yeats, R., Reddy, P. J., Wheeler, A., Senior, C., & Murray, J. (2010). What a difference a writing centre makes: a small scale study. *Education & Training, 52*(6/7), 499-507. doi: 10.1108/00400911011068450

Yorke, M., Barnett, G., Evanson, P., Haines, C., Jenkins, D., Knight, P., . . . Woolf, H. (2005). Mining Institutional Datasets to Support Policy Making and Implementation. *Journal of Higher Education Policy and Management, 27*(2), 285-298.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8*, 307-325.