Implications and Challenges to Using Data Mining in Educational Research in the Canadian Context

Samira ElAtia
University of Alberta

Donald Ipperciel
University of Alberta

Ahmed Hammad
HMD, Project Management Services

Abstract

Canadian institutions of higher education are major players on the international arena for educating future generations and producing leaders around the world in various fields. In the last decade, Canadian universities have seen an influx in their incoming international students, who contribute over $3.5 billion to the Canadian economy (Madgett & Bélanger 2008, p. 195). Research in Canadian post-secondary institutions is booming, especially in education (SSHRC, 2011)—for the academic year 2010-2011, of the 12 subject areas, the total SSHRC funding for projects in education, ranked fourth, exceeding $27 million. All of these variables place Canadian higher education in a leading and strategic position in several educational research fields. One can imagine the wealth of knowledge about trends in higher education that could be revealed if the large amount of data generated by Canadian universities were systematically analyzed and handled using techniques such as data mining. However, not much can be achieved from the unharnessed knowledge accumulated on a daily basis, as the advancement of data mining research that would provide the ultimate tool to learn about trends and changes in Canadian institutions is often held back by inadequate data warehousing, as well as by privacy, confidentiality, and copyright regulations. In this paper, we engage in a critical discussion/analysis of the interface between data mining research in higher education and the legal implications of such a tool.

CANADIAN JOURNAL OF EDUCATION 35, 2 (2012): 101-119

Résumé

Les établissements canadiens d'enseignement supérieur jouent un rôle majeur sur la scène internationale dans l'éducation des générations futures et dans la formation de leaders dans divers domaines à travers le monde. Au cours de la dernière décennie, les universités canadiennes ont connu un afflux d'étudiants internationaux, qui contribuent plus de 3,5 milliards de dollars à l'économie canadienne (Bélanger & Madgett 2008, p. 195). La recherche dans les institutions canadiennes d'enseignement postsecondaire est en plein essor, en particulier en matière d'éducation (CRSH, 2011) - pour l'année académique 2010-2011, parmi les 12 domaines de recherche, le financement total du CRSH pour les projets portant sur l'éducation, au quatrième rang, s'élevait à plus de 27 millions de dollars. Toutes ces variables placent l'enseignement supérieur canadien dans une position stratégique et de premier plan dans plusieurs domaines de recherche en éducation. On peut imaginer la richesse des informations sur les tendances dans l'enseignement supérieur qui pourrait être révélée si la masse de données générée par les universités canadiennes était systématiquement analysée et traitée en utilisant des techniques telle que l'exploration de données. Cependant, on ne peut guère obtenir grand chose à partir des informations accumulées sur une base quotidienne, étant donné que l'avancement de la recherche à exploration de données, qui serait l'outil ultime pour en apprendre davantage sur les tendances et les changements dans les institutions canadiennes, est souvent freinée par un entreposage de données insuffisant, ainsi que par les règlementations sur la protection des renseignements personnels, la confidentialité et le droit d'auteur. Dans cet article, nous engageons une discussion et une analyse critiques de l'interface entre la recherche à exploration de données dans l'enseignement supérieur et les implications juridiques d'un tel outil.

Implications and Challenges to Using Data Mining in Educational Research in the Canadian Context

Introduction

According to *The Economist* (2011), based on an EMC/IDC Go-to-the-Market study, there were 130 exabytes[1] of information generated in 2005; this number is forecasted to increase to 2,720 exabytes in the year 2012 and be triple that amount in 2015, at which point it is predicted to reach 7,910 exabytes (EMC/IDC 2011). Data are generated daily on every aspect of our lives. If used and analyzed properly, even though this data "will flood the planet […] it will help us understand it better" (Big data, 2011).

In institutions of higher education, the trend of growing amounts of data continues. For Siemens and Long (2011), "the most dramatic factor shaping the future of higher education is something that we can't actually touch or see: *big data and analytics*" (p.1). Large amounts of data from a variety of sources are collected daily on classes, students, administration, faculty members, programs of study, etc. Most of this generated data goes unprocessed. The little that is processed is confined to a specific inquiry or targeted research question. None of it is looked at from a 'big picture' perspective that combines all that is collected. Data are not inter-linked and are independent from each other. As a result, potentially important and valuable information is lost. Data is not stored nor treated as a single large entity in which more variables could be included and trends revealed.

The main cause for this situation and the loss of valuable information is the lack of a systematic approach for collecting, storing, codifying, and analyzing this data. This data, in the majority of cases, is initially not collected nor coded properly. It is stored away in formats that do not allow much analysis or extraction of useful knowledge. It is analyzed at the level of smaller units in which only interested parties can take advantage of it and, most importantly, in which research questions and variables are already pre-determined. Moreover, this data is completely unconnected and is stored in ways that do not allow any relationships to be built or discerning trends to be recognized.

Yet, if the same data were available to a larger research audience—in a format accessible to all, as well as being stored and coded in an integrated way so as to allow diverse academic users to access it, add to it, and analyze it according to their own perspectives—a wealth of knowledge could be harvested from this scattered data. In the current situation, different departments and units within universities collect and store pertinent data in different formats throughout the academic year. The procedure is time-consuming, and often costly. It is a huge loss for educational purposes that very large sets of collected data are hardly analyzed and are not transferred to useful knowledge that could be taken advantage of to address challenges in educational research for the 21st century.

In this context, this article aims to address the following questions. Regardless of the uniqueness of each program within the university,

1. Would it be feasible to develop an integrated data acquisition system for collecting and storing data from all departments and units within a university?

[1] 1 billion gigabytes.

2. Can the collected data be converted to useful knowledge and provide new insights into educational research?

3. Can such practices be done in a way that does not infringe on legal issues relating to privacy and confidentiality?

Defining Data Mining and Knowledge Discovery in Data models

Knowledge discovery in data. In a multifaceted environment in which data comes from different sources and in different shapes, the concepts of Knowledge Discovery in Data (KDD), data warehousing and data mining offer an alternative for learning from this data. These techniques are eclectic in nature and combine qualitative as well as quantitative research approaches; they also allow researchers to work with large amounts of data that are impacted by a large number of unknown variables. The KDD hybrid model in Figure 1 was adapted from Cios, Pedrycz, Swiniarski, & Kurgan (2007). The model starts by understanding the problem domain. The second step is to analyze the problem data and its multiple dimensions using targeted datasets. The next step consists of the development of the data collection model and prototype data warehouse. The data warehouse stores collected data in a ready-for-mining format that can be used for dynamic On Line Analytical Processing (OLAP) reports and graphs. This step is followed by the implementation of data mining techniques in case studies utilizing a large set of actual education data from an institution of higher education.

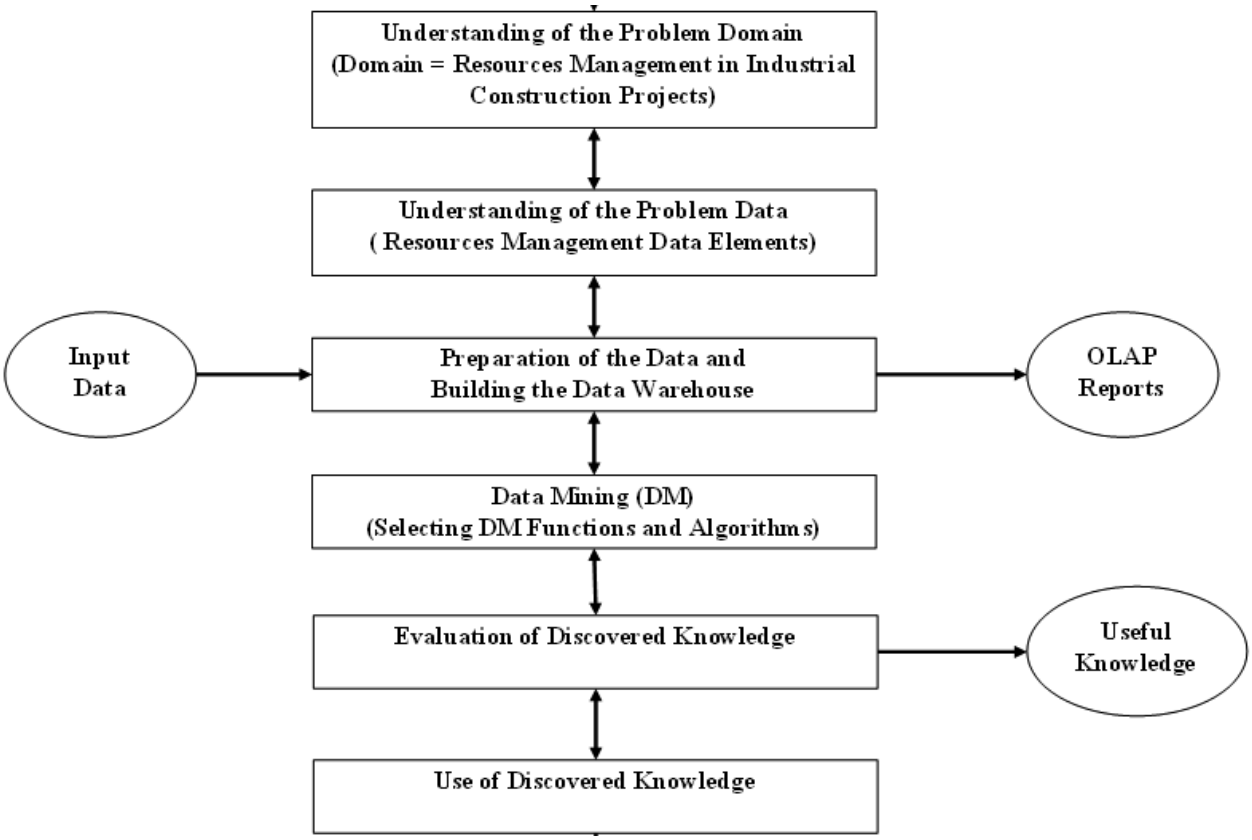

Figure 1. The Modified Hybrid KDD Model (Hammad, 2009).

One must point out that knowledge comes in different forms. This can be illustrated with the help of the knowledge pyramid (Liebowitz & Megbolugbe, 2003). It consists of data at its base, which leads to information, then knowledge, and finally wisdom at the tip of the pyramid, as shown in Figure 2. Data represents raw elements, and when these elements are patterned in a certain way, data is transformed into information. Once certain rules or "heuristics" are applied to this information, knowledge is then generated as actionable information for producing value-added benefits. Wisdom in the knowledge pyramid represents the ability to make knowledge-based decisions in order to maximize the value-added benefits.

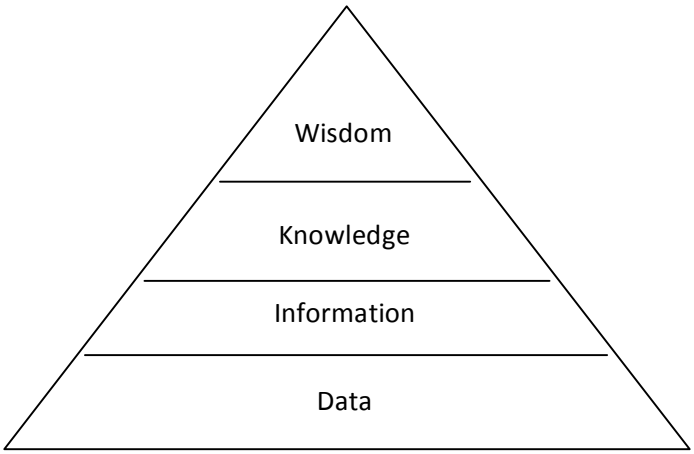

Figure 2. The Knowledge Pyramid.

Knowledge is typically classified into two categories: tacit and explicit. Explicit knowledge can be presented in the form of books, research papers, reports, graphs, memos, and drawings, whereas tacit knowledge is mainly stored in human minds and represents their experience. According to Davenport and Prusak (1998), knowledge is always generated in organizations while running any kind of project or doing any type of task or business. This knowledge needs to be codified to become accessible to those who need it. Once the codification process is complete, knowledge can then be transferred to others who, in return, can also generate new knowledge. These three steps that form the complete knowledge cycle are the backbone of the KDD model. These steps are also cyclic in nature and are part of a constant loop of feedback that modifies and improves the knowledge harnessed as more data is collected and analyzed.

Data warehousing. Data warehousing and data mining are amongst the techniques used to convert data into useful knowledge. Data warehouses are dedicated, read-only, and non-volatile databases that centrally store validated multidimensional historical data from Operation Support Systems (OSS) to be used by Decision Support Systems (DSS) (Inmon, 2005). The structure of data warehouses relies mostly on the star schema for simple datasets and on the snowflake schema for complicated datasets. The star schema consists of a fact table that contains data and dimension tables, which in turn contain the attributes of this data. The snowflake schema is used either when multiple fact tables are needed or when dimension tables are hierarchical in nature (Giovinazzo, 2000). According to Ahmad & Azhar (2004), a data warehouse typically comprises three main components: the data acquisition systems, also known

as back-end; the central database; and the knowledge extraction tools, known as the front-end. In the back-end of a data warehouse, the data can be extracted from text files, spreadsheets, or On Line Transactional Processing (OLTP) operational databases. In the front-end, On Line Analytical Processing (OLAP) techniques consolidate and aggregate datasets using variable paths to enable the dynamic data analysis according to decision-makers' needs (Codd, Codd,& Salley, 1993). According to Fan (2007), OLAP techniques include: roll-up and drill-down, slice and dice, and data pivoting. Roll-up and drill-down techniques are used to view data at different levels of details according to user's needs. Slice and dice are used to view the data either from one dimension or multiple dimensions. Data pivoting shows the data on a two-dimension matrix with multiple row and column headings. This technique also provides users with powerful graphical illustrations and filtering capabilities as per user's needs.

Data mining. Han and Kamber (2006) define data mining as "[t]he analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owners." Likewise, Fayyad, Piatetsky-Shapiro, and Smyth (1996) emphasize the fact that the discovered knowledge that ensues from the data mining procedure has to be previously unknown, non-trivial, and genuinely useful to the data owners. Data mining techniques rely on either supervised or unsupervised learning and can be grouped into four categories: clustering, finding association rules, classification and outlier analysis (Cios et al., 2007).

Clustering methods rely mainly on the concept of minimizing the distances between data points falling in a cluster and maximizing the distances between these data points and the data points in other clusters (Zaiane, 2002). Finding association rules is used to detect hidden patterns in large datasets. Classification techniques build a model using a training dataset to define data classes, evaluate the model and then use the developed model to classify each new data point into the appropriate class. Classification techniques include decision trees, rule-based algorithms, artificial neural networks (ANN), k-nearest neighbours (k-NN or lazy learning), support vector machine (SVM) and other data classification techniques (Cios, 2007).

Potential of Data Mining in Educational Research

Data mining in education. The application of data mining in educational systems is an "iterative cycle of hypothesis formation, testing, and refinement (Romero & Ventura 2007, p.135." Figure 3 below illustrates how this cycle operates. As more data is generated, more information can be learned and more knowledge can be harvested. In this cyclic process, the learning environment in general is adapted, reformed, edited and restructured as more information become available and more connections between data points are established. Baker (2011) identifies five categories of uses of educational data mining research: prediction, clustering relationship mining, distillation for human judgment, and discovery with models.

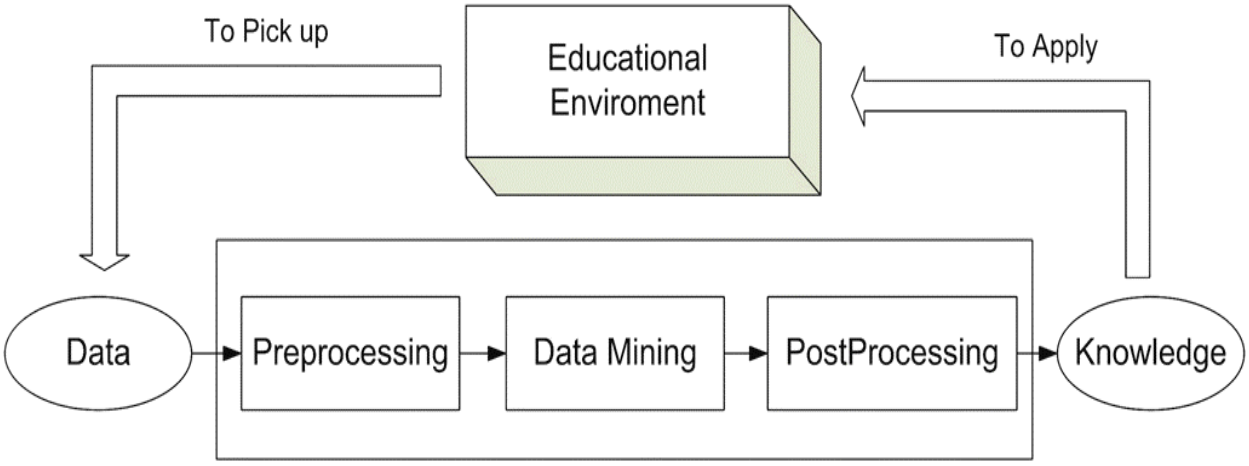

Figure 3. Educational data mining process (Garcia, Romero, Ventura, & de Castro, 2011)

Research using data mining techniques can be useful in three contexts:

a) when it is geared towards providing information that enhances learning from the students' perspective;

b) when it is geared towards providing insight into learning and teaching for university teachers; and

c) when it is geared towards providing the institution and its administrators with valuable knowledge that would be an asset for decision-making.

Gaudioso, Montero, Hernandez-del-Olmo (2012), using PDinamet as an adaptive educational system, found out that the data generated using this adaptive e-learning environment and these data mining techniques helped teachers in predictive and descriptive ways to address and target specific issues in students' learning. Chou, Wu, Li, & Chen (2009) used data mining to find ways to "[p]ersonalize curriculum sequencing" (p. 296) in a web-based learning environment under the assumption that no fixed learning path is appropriate for all learners. Relying on the e-learners' previous course records and navigation patterns in the web-based courses, they were able to predict behaviours and recommend proper approaches for successful learning experiences. In the same line of work, Wang, Tseng, & Liao (2009) used data mining techniques in order "to adapt course content to the learning needs of individual students" in the teaching of English as second language for Taiwanese students. Data mining was beneficial in decision-making for optimizing learning sequences and personalizing the learning.

The last five years saw an explosion of research targeting various educational/learning contexts. The educational data mining conferences in Canada and around the world are profuse, with research that uses data mining in innovative ways within classroom settings. For instance, Chen et al. (2009) used data mining in an educational social networking analysis, whereas Rabbany, Takaffoli, & Zaïane (2011) studied the interactions between students within an e-class environment for the purpose of finding a valid and fair evaluation model to assess this interaction using data mining techniques. They have built their model on previous work by Marin and Wellman (2010).

Martinez, Yacef, Kay, Al-Qaraghuli, & Kharrufa (2011) used data mining techniques to analyze group interactions in collaborative educational context where they focus not only on the end product but on the process of the collaborative work itself. With their study, they open a new angle for using data mining in educational research. Within a collaborative learning framework, Kharrufa, Leat, & Olivier (2010) designed a tool for the development and assessment of students' higher level thinking skills, a model that relies on data mining. Martinez et al. (2011) concluded that various data mining techniques can be successfully used and implemented to gain insight within a collaborative learning framework and environment and in group interactions, yielding positive results for working in collaboration (Anaya & Boticario, 2011), for problem solving and conflict resolution (Prata et al. 2009), for effectively working in teams (Perera, Kay, Koprinska, Yacef, & Zaiane,2009), and for the correctness and the appropriateness of the target learning tasks (Talavera & Gaudioso 2004).

Garcia et al. (2011) used data mining techniques in order to systematically provide teachers with feedback that would improve their e-learning courses. In their study, they admit that their model has only been used with teachers and experts who were "involved in the development of the tool itself." They hoped that in the future, they could generalize it to include other teachers who are not necessarily savvy with data mining algorithms and techniques. In the same line of facilitating the use of data mining for the general educational public, Zengin, Esgi, Erginer, & Aksoy (2011) conducted a study in which they were able to successfully implement several data mining techniques in an educational sciences environment.

Using special data mining technique, Lee (2007) was able to develop a model that "is reusable for diagnostic, predictive and compositional modeling in a fast and easily scalable manner" within an e-learning web-based environment. First, they were able to provide a *diagnostic modeling* that identifies the presence or absence of a certain action/behaviour among students. Based on this diagnostic modeling, the *predictive modeling* was constructed focusing on comparing the student behaviour with similar past student behaviour. The purpose of such a modeling tool is to predict what the student will do in order to make appropriate recommendations on how the students should proceed. Finally, the *compositional student modeling* allows reuse of the student model parameters in this iterative process.

The potential uses and applications of data mining are larger in number than what is stated above. From reviewing studies on the application of data mining in education, we notice that (1) all are centered on e-learning and the e-classroom in a computer-assisted learning environment, and (2) most of the research conducted is carried out by computer scientists who are savvy with the algorithms necessary for developing data mining techniques. There is hardly any research from educational scientists using DM in a KDD model, nor is there any research that focuses on applications for improving higher education using the data accumulated in most institution of higher education in a holistic way. However, Learning Analytics emerged in the last three years as field in its own right, one that deals specifically with this issue (i.e., bridging between data mining and the general educational environment).

Learning analytics and knowledge. The unique hierarchical features of educational data (Baker, 2011) provide researchers in educational environment with opportunities to use data mining for investigation. Stemming from and building on the field of data mining in educational settings, Learning Analytics and Knowledge (LAK) focuses on collecting, measuring, and analyzing data about learners and their learning contexts and environment for the purpose of optimizing these learning contexts. LAK bridges between the micro level of data mining to the macro level of educational research (Siemens & Baker, 2012) and aims to understand the learning environment. In 2011, the first international LAK conference took place in Banff,

Canada. The proceedings[2] of the conference offer a variety of research papers on LAK using data mining in educational contexts.

According to Zaïane (2012):

> Learning Analytics is data analytics in the context of learning and education; that is, the collection of data about learners' activities and behaviour as well as data about the environment and context in which the learning took place; and the analysis of such data using statistics and data mining techniques with the purpose of extracting relevant patterns from this data to better understand the learning that took place. The objectives of Learning Analytics can either be the reporting of measures and patterns from the data for the understanding of learning activities, or the optimization of the learning activities and strategies or the environments in which the learning occurs. (2012, Personal communication)

Needless to say, the potential of this emerging field is immense. Elias (2011) summarized the different cycles of LAK in Figure 4 below.

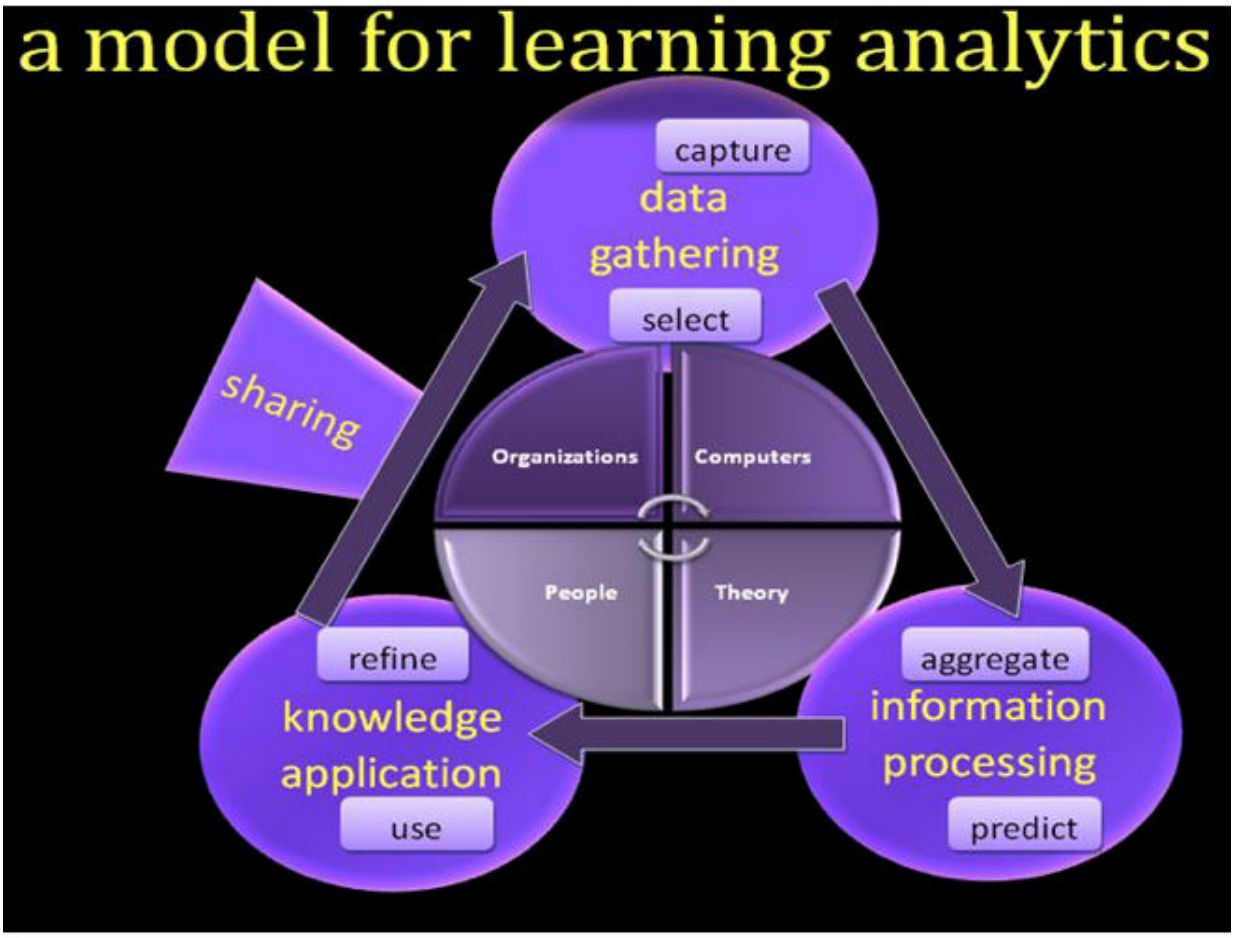

Figure 4. Learning analytics continuous improvement cycle (Elias 2011)

[2] http://dl.acm.org/citation.cfm?id=2090116&picked=prox&cfid=104612425&cftoken=34139900

Bienkowski, Feng, & Means (2012) provide a link between data mining and LAK whereby "*educational data mining* is looking for new patterns in data and developing new algorithms and/or new models, while *learning analytics* is applying known predictive models in instructional systems" (Bienkowski et al. 2012, p.8). On the one hand, data mining in educational environments focuses on automated responses to students; on the other hand, LAK "enables tailoring of responses, such as through adapting instructional content, intervening with at-risk students, and providing feedback" (Bienkowski et al. 2012, p.12). LAK uses data and meta data from data mining techniques that would enable the use and application of already established DM models to tackle various questions affecting students' learning within the institutional/organizational learning systems (Johnson et al., 2011).

The set of educational projects that could take advantage of data mining techniques are limitless. For instance, building on Gaudioso et al. (2012), one could devise an adaptive e-learning environment in which content is tailored not only to the learner's previous knowledge (cognitive dimension) but also to his/her interests (affective dimension). An early alert algorithm could identify at-risk students even before they run into actual academic difficulties. The relative difficulty of specific exercises could be ascertained, providing the instructor with valuable information and clues as to what issues should first be raised in class. The evolution of entire programs could be analyzed, thus giving academic administrators key insight as to how to improve on them, if need be. Along this line, ElAtia and Ipperciel (2011) investigated the possibility of using DM to analyze a longitude set of data spanning from 2000 to 2007, with information about evaluation procedures that post-secondary instructors use in assessing the students' achievement throughout the semester. The possibilities are indeed endless in all sectors of education. However, any attempt at using data mining in education must take into consideration strict legislation that regulates confidentiality and privacy issues.

Ethical and Legal Considerations for Educational Research

The issue of privacy as it relates to data mining is more prevalent in situations involving business dealings, targeted advertising, banking, social networking, internet searches, genetic material, and the like. In education, this issue has not been much debated, although it does raise questions with regard to its ethicality and legality. The use of technology, of data mining techniques in particular, can be at time intrusive. In fact, according to Nancy Holmes (2008), privacy issues historically arose in Canada in the late 1960's when computers were first being used on a large scale by government and large corporations. Even with the best of intentions, as in algorithms attempting to predict future failure or success in students, private information can be improperly gathered. Even in those cases, as in others, the collection of data is subject to clearly defined rules.

In Canada, the protection of privacy and private information is ensured by various legal acts, starting with the Canadian Charter of Rights and Freedoms, articles 7 and 8, which guarantee the right to security of the person, including the right to be secure against unreasonable search and seizure. The Charter came into effect in 1982. At that point, many countries—European for the most part—had enacted access to information acts. It is in the wake of this global movement, to which we will be coming back in a moment, that Canada's Privacy Act and Access to Information Act were passed in 1983. These were Canada's first pieces of legislation that focused specifically on privacy and access to information rights. The importance of these acts for Canadian political life is signified by the fact that it made provision for a Privacy Commissioner and an Information Commissioner who can, among other rights, investigate complaints and pursue court action.

These acts, however, are rather limited in scope, as their purpose is to regulate the actions of federal government institutions pertaining to the collection and disclosure of personal information in the case of the Privacy Act, and to citizens' access to government information in the case of the Information Act. After an expansion of these laws in 1996 with the Freedom of Information Act, which gave individuals the additional right to request correction of personal information about themselves, the most significant development in Canadian data privacy laws came with the Personal Information Protection and Electronic Document Act (PIPEDA), enacted in 2000. This law applies specifically to private sector organization, regulating collection, use, and disclosure of personal information. Once more, this legislation came about in response to pressure from the European Union, which was concerned about the personal information of its citizens as they interacted with Canadian institutions and businesses.

Four Canadian provinces have equivalent legislation: Quebec, British Columbia, Alberta, and Ontario. The provinces' acts are somewhat more restrictive than their federal counterpart. They are particularly interesting in our academic context insofar as they include universities as "organizations" covered by the law. For instance, Alberta's information privacy act, known as the Freedom of Information and Protection of Privacy (FOIP) was passed in 1995, with post-secondary institutions being added to the list of public bodies covered in 1999. Any collection, use, or disclosure of personal information in the context of university research is subject to FOIP legislation. Most Alberta universities have a FOIP office and/or coordinator that oversees data privacy matters and educates the university population about data protection and privacy issues.

The Canadian laws on privacy are part of a global system of rights that are well entrenched in international law. The Universal Declaration of Human Rights, adopted in 1948 by the United Nations General Assembly, formulates a privacy right in article 12: "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." Similar provisions are also provided in the International Covenant on Civil and Political Rights, article 17, adopted in 1966 (United Nations 1966). Europe, for historical reasons pertaining to events during the Second World War, has been a leader in promoting and adopting privacy laws. The European Convention on Human Rights, adopted in 1950, affirms in article 8 a right to privacy to which all member states are subject (Council of Europe, 1950). And the American Convention on Human Rights, adopted by 24 of the 35 Organization of American States (OAS) member states in the 1970's and 1980's,[3] also stipulates similar rights protecting private life (see article 11; OAS, 1969). Outside of Europe and the Americas, privacy rights were affirmed by 21 APEC members in its "Privacy Framework" (2005).

With a view to harmonize privacy laws and practices among Organisation for Economic Co-operation and Development (OECD)member states, Guidelines for the Protection of Privacy and Transborder Flows of Personal Data were produced in 1980 (OECD, 1980), Canada being a signatory since 1984. The Canadian Standards Association's Model Code for the Protection of Personal Information, stating 10 principles (CSA, 1996), is based on these Guidelines,[4] and so is the Fair Information Practice Principles produced by the US Federal Trade Commission (FTC, 1998), although the latter has been criticized for coming short of the OECD Guidelines. In fact, the OECD Guidelines were the basis for data protection legislation in many countries, including Canada, and for policy used by non-government organization. Their interest lies in their stating

[3] Canada is not a signatory of this treaty mainly because of its pro-life provisions (article 4.1).
[4] See http://www.csa.ca/cm/ca/en/privacy-code/publications/view-privacy-code/article/introduction

explicit principles that should be followed in order to ensure adequate data protection. The Guidelines identify eight such principles: (1) Collection Limitation, (2) Data Quality, (3) Purpose Specification, (4) Use Limitation, (5) Security Safeguards, (6) Openness, (7) Individual Participation, and (8) Accountability. For our purpose, as Canadian scholars, we would like to consider the similar principles identified by the CSA Model Code, that is, (1) Accountability; (2) Identifying Purposes; (3) Consent; (4) Limiting Collection; (5) Limiting use, disclosure, and retention; (6) Accuracy; (7) Safeguards; (8) Openness; (9) Individual Access; (10) and Challenging Compliance. These are considered minimum requirements for the protection of personal information.

1. Accountability. The institution in which we work is ultimately accountable for personal information. That is why a person or an office is usually designated in post-secondary institutions to see to it that privacy rules are complied with. Considering its novelty in educational research, any data mining project should be discussed with the institution's privacy officer (or any designated individual responsible for data protection).

2. Identifying purposes. Prior to any data collection, the purpose for which the information is collected must be identified and the use of data must be restricted to the stated purpose. This can sometimes pose some problems in data mining projects, as data mining often extracts unforeseeable patterns from data sources and may lead to new purposes being defined. In those cases, new purposes can indeed be defined but only subsequent collections of data can be used for this new purpose.

3. Consent. Informed consent of individuals for the collection, use, and disclosure of their personal information within the frame of the identified purpose is required whenever possible. When working with large data sets including, for instance, tens of thousands of students, this is obviously impractical. This is why one must make sure that in this case information is "depersonalized"(i.e., made impossible to be traced back to specific individuals). To this effect, many privacy-preserving techniques are used in data mining (see for example, Charu & Philip 2008). One should keep in mind that individuals have the right to withdraw consent at any time.

4. Limiting collection. Good practice in personal data management requires that personal information is not collected indiscriminately but rather limited to what is necessary given the scope of the identified purpose. However, data mining is precisely a form of knowledge discovery in which, as we have stated earlier, "unsuspected relationships" are found. In this context, it is not known beforehand which data is relevant. This is why large databases containing indiscriminate data are often collected in data warehouses. Again, this principle can be satisfied by depersonalizing the collected information.

5. Limiting use, disclosure, and retention. According to this principle, personal information should be used, disclosed, and retained strictly for the identified purpose. It follows from principles 2, 3, and 4 and face similar difficulties. Here, as in principle 3 and 4, the solution to this difficulty is producing depersonalized data.

6. Accuracy. Personal information should be as accurate and up-to-date as possible, given the stated purpose, as the collected information will be used to make decisions about

individuals. This principle has as much and operational dimension—the efficiency of the project depending on accurate information—as an ethical one, as potentially adverse decisions can be minimized with accurate information. If, for instance, at-risk student are not identified because of inaccurate information, avoidable adverse consequences will certainly ensue.

7. Safeguards. Personal information should be treated as highly sensitive data that needs to be adequately protected against unauthorized access, modification, theft, etc. Security safeguards include physical measures (e.g., restricted access to offices or file cabinets containing the personal information), administrative measures (e.g., security clearances) and technological measures (use of passwords and encryption). Following this principle, data warehouses should be set in the institutions' secured data center, in a password-protected server.

8. Openness. The duty of openness requires that policy and practice information pertaining to personal information management is made available to all individuals who wish to access it. This requirement is normally taken on by the institution's privacy office and any complaint or inquiry should be directed toward this office. However, the researcher's contact information should also be included, as a person who is accountable for the management of personal information in the context of a specific data mining project.

9. Individual access. Individuals have the right, upon request, to be informed of the existence, use and disclosure of their personal information and they have the right to request that inaccurate or incomplete information be corrected. Often in data mining projects, information is depersonalized and cannot, in this context, be linked to specific individuals or be corrected in cases of inaccuracies. However, when data is indeed linked to specific individuals, as it is the case in the example of early detection of at-risk students, individuals are entitled to know of its existence and to correct it if need be.

10. Challenging compliance. This principle expresses nothing other than a right to lodge a complaint with the accountable people or entity (i.e., the researcher or the institution) if any of the stated principles are not respected. In this case, one must normally follow the procedure put in place by the institution.

As one can see, constraints on data collection, warehousing, and use are many and can pose a challenge to researchers wishing to conduct data mining projects. Data mining can advance Canadian Educational research and put it at the forefront of global innovation in research. Yet, the very nature of data mining research requires an open access to data, since the presupposition of data mining is that the researcher does not target a specific variable(s) and follows a certain rigid methodology and statistical procedures to answer a specifically formulated research question(s): in data mining, we observe, and let the data reveal itself. Data mining allows researchers and other stakeholders to have an overarching look at the data, its working, and its underlying relationships.

Conclusion

In the global context, Canadian institutions of higher education are well-equipped to take the lead in using data mining resources for educational research and to improve the education experience on all levels. These levels can be grouped into four major categories:

1. Focus on the learning: In this context, using data mining can be helpful on different levels: from a macro-level of assessing students' academic progress for retention and for addressing potential problems ahead of time, to a micro-level of ensuring a better learning experience and an involvement of the students in taking control and being actively engaged in their own learning.

2. Focus on the teaching: The use of data mining by academic staff would greatly enhance the teaching approaches. Teachers can use data mining to find innovative ways to help improve learning and teaching as well as develop assessment procedures for fairer and more valid results within a formative evaluation framework. Data mining can also be a valuable tool for teachers who seek formative feedback for their instruction.

3. Managing the institution from an academic experience: From the institution's point of view, different units within the university—e.g., the library, the registrar's office, the IT department, students' affairs, etc.—can benefit from using data mining in following the students' progress. When data from these units is made accessible for data mining, new relationships and variables can emerge, which could be very useful with regards to student retention, assisting at-risk students well in advance of any actual problems, and ensuring that the academic progress is smooth.

4. Managing the institution from an infrastructure perspective: With the budget crunches that post-secondary institutions have been facing in the current economic recession, looking at the big picture through the wealth of data integrated from all departments and units in the institution, data mining techniques would offer a unique opportunity to show new connections and new insights. In so doing, the working of a whole institution can be studied globally instead of unit by unit, each conducting their own research independently and trying to find ways to optimize their working environment.

With data mining, we do not know what connections we will discover—it is truly a tool that lets the data speak without any manipulation from outside forces, particularly personal views and biases of the researchers. Data can reveal patterns that no one could have foreseen and, most importantly, it takes into consideration each element in the database, each being unique and handled according to the supplied information. With data mining in educational research, Canadian institutions of higher education can be at the forefront of cutting edge advancement in research, but the intentions of the university stakeholders need to be made clear with regards to fair use of data in a way that does not hinder the confidentiality and privacy of all parties linked to the institution. As a result, applying data mining at a broad institutional level may prove to be a greater challenge from an administrative rather than a technical perspective.

At the core of any research project that aims to advance education using data mining techniques stands the right of the individual for privacy and confidentiality. The connections that

one seeks to achieve in such projects necessitate, technically and conceptually speaking, that various data 'speak to each other,' for all parties involved. In itself, such a premise raises red flags with regard to proper handling of data wherever the principles of confidentiality, anonymity, privacy and the ethos of fair professional conduct are ingrained and respected. So far, most models of research in data mining have more of less been conducted on a nuclear level within a classroom or a set of students. However, as we see the potential of data mining techniques within a LAK framework expanding to increasingly large-scale operations, educational institutions in conjunction with policy-makers must turn their attention to the pressing matter of defining and ensuring proper and fair use of personal data in education research using data mining in a manner that does not hinder scientific progress. As the fields of educational data mining and LAK advances in high gear, the educational community needs to catch up for its own sake.

References

Ahmad, I. & Azhar, S. (2004). Data Warehousing in Construction: From Conception to Application. In S. J. Krishna (Eds.), *Data Warehousing: Design and Development Perspectives*. ICFAI Books, India.

Anaya, A.R.,& Boticario, J.G. (2011). Application of machine learning techniques to analyse student interactions and improve the collaboration process. *Expert Systems with Applications, 38*, 1171-1181.

APEC – Asia-Pacific Economic Cooperation (2005). *Privacy framework*. Singapore: APEC Secretariat. Retrieved from http://www.ag.gov.au/www/agd/rwpattach.nsf/VAP/%2803995EABC73F94816C2AF4AA2645824B%29~APEC+Privacy+Framework.pdf/$file/APEC+Privacy+Framework.pdf.

Baker, R. S. J. d. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International Encyclopedia of Education* (3rd ed., vol. 7, pp. 112–118). Oxford, UK: Elsevier.

Baker, R.S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*, 1, 3-17.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Office of Educational Technology, U.S. Department of Education. Retrieved from http://evidenceframework.org/wp-content/uploads/2012/04/EDM-LA-Brief-Draft_4_10_12c.pdf

Big data: Drowning in numbers. *The Economist*. Retrieved from http://www.the economist.com

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, *13*, 156–169.

Charu, A., & Philip, S. Y. (Eds.). (2008). *Privacy-preserving data mining: Models and algorithms*. New York. Springer.

Chen, J., Fagnan, J., Goebel,R., Rabbany, R., Sangi, S., Takaffoli, M, Verbeek, E., & Zaiane, O.(2010). Meerkat: Community mining with dynamic social networks. In *Data Mining Workshops (ICDMW)* (pp. 1377-1380), 2010 IEEE International Conference.

Chou, P.H., Wu, M. J., Li, P. H., & Chen, K. K. (2009) Accessing e-learners' knowledge for personalization in e-learning environment. *Journal of Research and Practice in Information Technology*, *41*(4), 295-318.

Cios, K., Pedrycz, W., Swiniarski, R. & Kurgan, L. (2007). *Data mining: A knowledge discovery approach.* New York: Springer.

Codd, E. F., Codd, S. B., & Salley, C. T. (1993). *Providing OLAP (On-line Analytical Processing) to user-analysts: An IT mandate.* San Jose, CA, USA: Codd and Date.

Council of Europe (1950). *The European convention on human rights*. Retrieved from http://www.hri.org/docs/ECHR50.html.

CSA – Canadian Standards Association (1996). *Model code for the protection of personal information*. Retrieved from http://www.csa.ca/cm/ca/en/privacy-code/publications/view-privacy-code/article/principles-in-summary.

Davenport, T. H., &Prusak, L. (2000). *Working knowledge: How organizations manage what they know.* Boston, MA: Harvard Business School Press, c1998.

D'mello, S., Olney, A.,& Person, N. (2011). Mining collaborative patterns in tutorial dialogues. *JEDM - Journal of Educational Data Mining, 2*, 1-37.

ElAtia, S., Ipperciel, D., & Hammad, A. (2011). *Advancing educational research through a Knowledge Discovery in Data (KDD) Model.* A paper presented at the American Educational Research Association annual convention at New Orleans, USA.

Elias, T. (2011). *Learning analytics: Definitions, processes and potential*. Retrieved from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf.

EMC/IDC Digital Universe Study. (2011). Retrieved from http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm.

Fan, H. (2007). *Leveraging operational data for intelligent decision support in construction equipment management*.(Doctoral dissertation, University of Alberta, Canada).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37.

FTC – Federal Trade Commission (1998). *Fair information practice principles*. Retrieved from http://www.ftc.gov/reports/privacy3/fairinfo.shtm.

Garcia, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *Internet and Higher Education*, *14*, 77–88.

Gaudioso, E., Montero, M., & Hernandez-del-Olmo, F. (2012). Supporting teachers in adaptive educational systems through predictive models: A proof of concept. *Expert Systems with Applications, 39*, 621-625.

Giovinazzo, W. A. (2000). *Object-oriented data warehouse design: Building a star schema.* Prentice Hall, Upper Saddle River, NJ.

Hammad, A.M. (2009). An integrated framework for managing labour resources data in industrial construction projects: A Knowledge Discovery in Data (KDD) approach. (Unpublished doctoral dissertation, University of Alberta, Canada).

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques.* San Francisco, CA: Morgan Kaufmann.

Holmes, Nancy (2008). *Canada's federal privacy laws.* Ottawa: Library of Parliament, Parliamentary Information and Research Service. Retrieved from http://www.parl.gc.ca/Content/LOP/researchpublications/prb0744-e.pdf

Inmon, W. H. (2005). *Building the data warehouse.* Indianapolis, IN: Wiley.

Johnson, L., Smith, R., Willis, H., Levine, A. & Haywood, K. (2011). *The 2011 Horizon report*. Austin, Texas: The New Media Consortium. Retrieved from http://www.nmc.org/pdf/2011-Horizon-Report.pdf

Kharrufa, A., Leat, D., & Olivier, P. (2010). Digital mysteries: Designing for learning at the tabletop. In *ACM International Conference on Interactive Tabletops and Surfaces*. Saarbrücken, Germany: ACM, 197-206.

Lee, Chien-Sing. (2007). Diagnostic, predictive and compositional modeling with data mining in integrated learning environments. *Computers and Education*, *49*, 562–580.

Liebowitz, J., & Megbolugbe, I. (2003). A set of frameworks to aid the project manager in conceptualizing and implementing knowledge management initiatives.*International Journal of Project Management, 21*(3), 189-198.

Madgett, P., & Bélanger, C. (2008). International students: The Canadian experience. *Tertiary Education and Management*, *14*(3), p. 191-207.

Martinez, A., Yacef, K., Kay, J., Al-Qaraghuli, A.,& Kharrufa, A. (2011). Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero , C., and Stamper, J. (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining, 13*(2), 111-120.

OAS – Organization of American States (1969). *American convention on human rights: "Pact of San Jose, Costa Rica."* Retrieved from http://www.oas.org/juridico/english/treaties/b-32.html.

OECD (1980). *Guidelines on the protection of privacy and transborder flows of personal data*. Retrieved from http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html.

Perera, D., Kay, J., Koprinska, I., Yacef, K.,& Zaiane, O. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Tran. on Knowledge and Data Engineering*, *21*, 759-772.

Prata, D., Baker, R., Costa, E., Rosé, C., Cui, Y.,& De Carvalho, A. (2009). Detecting and understanding the impact of cognitive and interpersonal conflict in computer supported collaborative learning environments. In *2nd International Conference On Educational Data Mining*, 131-140.

Rabbany R., Takaffoli, M., & Zaiane, O. (2011). Social network analysis and mining to support the assessment of online student participation. In Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero , C., and Stamper, J. (Eds.) *Proceedings of the 4th International Conference on Educational Data Mining, 13*(2), 20-31.

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

Siemens, G., & Baker, R. S. J. d.(2012). Learning analytics and educational datamining: Towards communication and collaboration. LAK12: International Conference on Learning Analytics & Knowledge, 29 April – 2 May, Vancouver, BC, Canada.

Siemens, G., & P. Long. (2011). Penetrating the fog: Analytics in learning and education.*EDUCAUSE Review*,*46* (5).

Social Sciences and Humanities Research Council of Canada. (2011**).** Canada's social sciences and humanities research community. *Facts and Figures.* Retrieved from http://www.sshrc-crsh.gc.ca.

\Talavera, L.,& Gaudioso, E. (2004). Mining Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces. In *Proceedings of the European Conference on Artificial Intelligence*. Presented at the 16th European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain.

United Nations (1966). *International covenant on civil and political rights*. Office of the United Nations High Commissioner for Human Rights. Retrieved from http://www2.ohchr.org/english/law/ccpr.htm.

Wang, Y.H, Tseng, M.H., & Liao, H.C (2009). Data mining for adaptive learning sequence in English language instruction."*Expert Systems with Applications*,*36*, 7681–7686.

Zaiane, O. R., Foss, A., Chi-Hoon Lee, & Wang, W. (2002). On data clustering analysis: Scalability, constraints, and validation.*Proceedings,* Springer-Verlag, Berlin, Germany, 28-39.

Zengin, K., Esgi, N., Erginer, E., & Aksoy, M.E. (2011). A sample study on applying data mining research techniques ineducational science: developing a more meaning of data. *Procedia Social and Behavioral Sciences*,15, 4028–4032.