Chapter 6

DATA ANALYTICS FOR EDUCATION AND HEALTHCARE

Maiga Chang

School of Computing and Information Systems, Athabasca University, Canada

- Abstract: Information systems always collect a lot of data for report and record purpose. The most common reports to be seen are showing users statistics based summary. In this chapter I would like to briefly share three of my research group's data analytics research and describe how they can be used in education and healthcare domain. The first research is an algorithm which can figure the difficulty of an online multiple choice question that most of students perceived while answering the question. The proposed algorithm can therefore give teachers feedback by identifying a student's learning situation as well as give reward students properly based on their performance improvement. The second research is Next Stop Recommender, a mobile app which can provide users recommendations for their next visit according to the similarity of their travel patterns between each other. Users have similar learning interests or difficulties can benefit from the system recommendations. The last research is a method that predicts the potential sitter requests a hospital may have in next week or even the very next day. The predicted results allow hospitals to adjust their strategies on resource assignments to better handle patient needs; for instance, they can allocate necessary resources like beds and medical professionals who have particular skills for the potential forthcoming patients in advance.
- Key words: Data Analytics, Mobile App, Data Mining, Frequent Pattern, Healthcare, Education

75

1. ONLINE QUIZ'S DIFFICULTY ANALYSIS

One area of concerns is that students are not prepared well when they graduate from high school (Alphonso, 2013). The 2012 International Student Assessment results from the Organization for Economic Co-operation and Development (OECD) show that Canadian scores in mathematics dropping significantly (Brochu, Deussing, Houme, & Chuy, 2013). Many research have been conducted to seek ways out to raise students' math skills and believe that technology can help.

Many Canadian provinces include Alberta have been studying on the benefits that students may have with the use of their own mobile devices in class and a large number of Canadian students have been used mobile devices in schools (Alberta Education, 2012; Joyce, 2014). Practi is educational software that has two components: *Press* and *Play*. *Press* is a web-based authoring tool that allows teachers to create educational content, instantly publish it to student mobile devices and monitor student progress. *Play* is an educational mobile app running on both of iOS and Android platform and engages students in meaningful and gamified skills practices through completing quizzes and tracking their own performance. Figure 1 shows the user interface of *Play* and the student has added four courses.



Figure 1. Student has four courses added to his or her Play.

Teachers can create many quizzes for a course and as many questions as possible for each quiz; for example, Figure 2(a) shows that the course "Math Fun" has four quizzes. At top right corner of Figure 2(b), there is an indicator "1/5" representing the student is working on the first question out of five that quiz "Math Fun 1" has.

Play aims to engage students in practicing more so they can be successful in particular academic subjects later; therefore, students are allowed to try on same quizzes and correspondent questions as many times as possible and *Play* records how much time the students spent on solving a question and traces their behaviours such as how many trials before they got correct answer for the question or skipped the question – gave the question up and turned to the next question in the quiz.



Figure 2. A course can have many quizzes and each quiz has many questions.

Figure 3 shows four students' behaviours of solving a question, #2212. Andy, Ben, Carl, and David tried to solve the question between November 15 and November 19. Andy and Carl skipped the question after several trials but Ben and David managed to get their answers right after several attempts (i.e., Ben) and found the correct answer quickly (i.e., David).

To get student motivated in terms of practicing repeatedly, *Play* needs to have a method that can reward them properly for practicing a question they are not familiar with or feel difficult from solving it earlier. Such method should be capable of calculating proper reward points for students who solve a question based on their question solving behaviours and the difficulty that most of students perceived while solving the question. The proposed method first categorizes students' question solving behaviours into eight patterns as Figure 4 shows.

- 4)



Figure 3. Four students' behaviours of solving question #2212.

		Correct or skipped (Correct = 4, Skipped = - Attempts state (H = -2, L = 2) Time state (H = -1, L = 1)				
Patterns		Difficulty	Equation	Weight	Symbol	
Correct and attempts (L) with time (L)		Easy	=4+2+1	7	А	
Correct and attempts (L) with time (H)		Easy	=4+2-1	5	В	
Correct and attempts (H) with time (L)	1 1 1	Easy	=4-2+1	3	С	
Correct and attempts (H) with time (H)		Normal	=4-2-1	1	D	
Skipped and attempts (L) with time (L)		Normal	=-4+2+1	-1	Е	
Skipped and attempts (L) with time (H)	L	Hard	=-4+2-1	-3	F	
Skipped and attempts (H) with time (L)	1	Hard	=-4-2+1	-5	G	
Skipped and attempts (H) with time (H)		Hard	=-4-2-1	-7	н	

Figure 4. Eight patterns of solving a question in Play.

The method believes that a student may feel the question is easy when he or she can correctly answer the questions with only few attempts OR in a short time – he or she had answered the question CORRECTLY when the student's attempt number is LOW OR the time he or she spent is considerably LESS. On the other hand, a student may have difficulty in solving the question if he or she skipped the question after many trials OR spent a lot of time – he or she SKIPPED the question while his or her attempt number is HIGH OR the time he or she spent is considerably HIGH. For other behaviours such as correctly answered question after many attempts and skipped the question quickly, the perception towards the difficulty of the question is categorized to normal.

In order to measure the perceived difficulty of a question that most of students may have, the method assigns weights for individual attribute values; for instances, it gives +4 for correctly answer and -4 for skipping the question; -2 for making many attempts and +2 for making only few attempts; and -1 for spending a lot of time and +1 for spending less time on the question. By summing up the weights, each pattern has its own weight from +7 to -7 and is represented by a symbol. At any given time frame, a student's behaviours of practicing a question can be treated as a pattern sequence like Figure 5 shows.



Figure 5. A pattern sequence of solving question #2212 during Nov. 15 to 18.

The second step that the method takes is to identify the frequent patterns from a dataset of sequences as Figure 6 shows. The frequent patterns can be seen as the most common behaviour pattern sequence that students have while solving the question and the weight of a frequent pattern can present how the difficulty of the question that a good portion of students perceived. The detailed process of identifying frequent patterns can be found at (Li, Kuo, Chang, & Garn, 2015). In the end of the process, the method finds three frequent patterns "A, A, B", "D, A, A", and "D, C, D" and their correspondent support values are 58%, 33%, and 33% – a pattern's support value means how many times the pattern can be found in the dataset, e.g., pattern "D, C, D" has been found in four sequences out of the twelve shown in Figure 6.

Question ID	User ID	Pattern sequence
2212	Andy	H, F, D, D, B, A, E
2212	Ben	D, C, E, D, C, D
2212	Carl	H, D, E, C, A, E
2212	David	A, A, B, E, C, E, E
2212	Anthony	A, C, A, A, B
2212	Derek	E, D, C, D, A, A, B
2212	Evan	F, D, A, A, B
2212	Bill	G, G, D, C, D, A, A, B
2212	Adam	F, G, C, A, A, B
2212	Edwin	F, F, D, D, B, B
2212	Denny	C, D, C, D, A, A, B
2212	Edgar	E, E, B, B, A, A, B

Figure 6. Sample dataset of pattern sequences for solving question #2212.

According to Figure 4, the weight that pattern "A" has is +7, pattern "B" has is +5, pattern "C" has is +3, and pattern "D" has is +1. The average weight of a frequent pattern then can be calculated out, i.e., "A, A, B" is 6.33, "D, A, A" is 5, and "D, C, D" is 1.67. The average weight tells us that both of patterns "A, A, B" and "D, A, A" can consider as easy pattern and the pattern "D, C, D" is between easy and normal pattern but is closer to normal pattern. The three frequent patterns show that the two easy patterns have total 91% support but the normal pattern has only 33% support. Under such circumstance, the question can be considered as an easy one based on the extracted common question solving patterns that most of students have while solving it. More detailed algorithm can be found at (Li, Kuo, Chang, and Garn, 2015).

The abovementioned data analytics method can also be adopted to analyse the usage data (i.e., feature usage, actions taken and browsing behaviours) that a system may record and store. The retrieved frequent pattern from the dataset of sequences can be treated as most common behaviours that users may take or have while using the system (e.g., learning management system, technology-enhanced learning environment, mobile learning app and educational game) or doing learning activities (e.g., solving problems and working on a worksheet).

2. NEXT STOP RECOMMENDER

When students freely learn in the real world environment, their wandering behaviours may represent their interests; for instance, a student may want to see lions and tigers first in a zoo while some others wanting to see butterflies first. On the other hand, if students are in a field trip with a worksheet, they may try to pay animals nearby a visit or go to see those animals they know where they could find. In the former case, the extracted wandering pattern for a student can be seen as his or her interests; on the other hand, in the latter case the extracted pattern can represent the student's problem solving behaviour. Liu and Chang (2011) propose a next-stop recommendation algorithm that can find top-N recommendations for users via the pattern extraction of users' wandering behaviours in real world and the match of two similar patterns. The algorithm can make a mobile learning system capable of offering students personalized learning service.

The proposed algorithm is briefly explained with the following scenario. Assuming there are five students learning in the real world and their learning routes can be told from Figure 7. The learning route of user #1 is recorded as $\langle J \ G \ C \ A \ E \rangle$ due to him or her had visited the Historical Building (J), Cultural Site (G), Person in History (C), another Cultural Site (A), and Statue V (E) one follows another.



Figure 7. Students' learning routes in the real world.

Now, the sixth student just left Cultural Site (A) and his or her recorded learning route is <J G C>, what next spot should the mobile learning system recommends him or her to go? Similar to the method of finding frequent patterns described in Section 1 earlier, several frequent patterns can be found

from the learning route dataset and the recommendation rules can be chose from the calculation results listed at top-left corner of Figure 8.

Because the student just left Cultural Site (A), rules #2, #3, #5, and #7 may not be good for the system to choose due to they have no spot A involved. Moreover, although rule #7 includes spot A, the rule has only two spots B and A and the student hasn't been spot B earlier so the rule should also be excluded due to the wandering pattern presented by the rule is very different from the student's behaviours. For more details, please refer to (Liu & Chang, 2011).



Figure 8. The workflow of finding eligible rules for making recommendations.

At right side of Figure 8, two tables includes rules #1, #4, and #6 shows the calculation of similarity between each of the rules and the student's wandering behaviours. The support and confidence that individual rules have based on the dataset of all students' learning routes. Taking rule #1 as example, the length of the student's recorded route is 4 (i.e., four spots, <J G C A>, have been visited) and rule #1's left-hand side includes two spots <G A>. Both of them had been paid a visit by the student, so the similarity between rule #1 and the student's route is 0.5. At the end, a bias weight is adopted by the recommendability equation as the bottom of Figure 8 shows. If a system believes that similarity is much important than the confidence and support a rule has, then it can set higher weight for similarity; on the other hand, it should set lower weight for similarity.

Figure 9 shows the mobile app implemented by Ripley et al. (2013) based on the above-mentioned algorithm. The mobile app can get its user's locations not only via built-in GPS reciever but also allow the user to manually enter the longitude and latitude where she or he is at as well as to scan a QR code that contains the coordinates with built-in camera as Figure 9(a) shows. When the user asks for, the app can try to find the top-N (e.g., top 2) recommendations for her or him. Figure 9(b) shows that two places have been recommended for the user to visit.

Next Stop Recommender	Next Stop Recommender
🚷 Logged in as: ben	🦣 Logged in as: ben
Station Svc: Running	Recommendations
Conver Connected	
Real Manual Entry	Southlake Regional Health Centre
QR Code Entry	Magna Centre
Get Recommendation	Cancel
Location Svc	
Logout	
Settings	
(a)	(b)

Figure 9. Next-Stop Recommender app.

The proposed algorithm can be implemented and integrated into any mobile learning systems to enable its ability of tracking students' learning behaviours in an authentic environment and making recommendations for students in terms of artefacts and objects that they might be interested or need in their learning process or field trip.

According to Figure 8, the system is capable of giving maximum three spots for the student to visit. Depending on the need of the system, the system can give the students two recommendation if the system chooses to only have top-2 rules found for the student or if the system has the recommendibility threshold set to 0.3. Figure 10 shows that teachers can check whether or not a student took the recommendation made by the system.

Chapter 6



Figure 10. A student took system' recommendation and paid a visit to the place.

3. MINING CLINICAL AND ADMINISTRATIVE HEALTHCARE DATA

In a healthcare setting such like hospital, a lot of information about patient are recorded in various systems (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Many reports can be generated for administrative staffs or managerial level personnel, but most of the reports have only counts, sums and groupings of collected data. Although some visualized ways can be used to the reports to facilitate data representation, they are mostly visual appeals or pivot tables that not necessarily provide more knowledge or discovery of new information for people.

Data mining methods can be used to discover patterns and relationships from a large dataset but the methods cannot tell users how important a discovered pattern is for them (Seifert, 2004). Many meaningful patterns can be analyzed and extracted from regular expression – a finite-state automata that is used for string pattern recognition (Jurafsky & Martin, 2000). It is also a way of describing complex patterns in texts. Regular expression has been used to extract information in biomedical field and provided an alternative approach to do complex semantic parsing (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001; Grishman, 1997; Mutalik, Deshpande, & Nadkarni, 2001). It uses shorter and simpler way to represent a long sequences that contains repeated patterns. For example, the sequence "AAABBBCCCABBCCCCC" can be represented by the regular expression "(A+)(B+)(C+)".

Many research have been done on transcribing and analyzing physicians' notes with regular expressions (Bhatia et al., 2010; Boudin, Nie, & Dawes, 2010; Chapman, Chu, & Dowling, 2007; Jia, Li, Dong, Long, 2011). Regular

expression techniques can be used to search patterns from data stored in string form and statistical methods could be then applied to find out particular string patterns. Some prediction analysis research have done with regular expressions; for instances, a proposed sequence analysis model can predict outpatient paths and patient flows (Konrad & Lawley, 2009) and a simple vector space model based diagnosis system that can extract knowledge from textbooks to provide readers hints about symptoms and possible treatments (Zhu, Fu, Xu, & Zhang, 2011).

A recommender system can be built based on the methods of pattern analysis and regular expressions to facilitate and improve the healthcare process, when it cooperates with patient care and clinical information systems. Most healthcare institutions collect huge amount of data in different systems. Each system contains valuable information but within a limited scope. It will be better to combine data from different data sources to perform data analysis, as the combined data contain information from different aspects and perspectives.

Sitter is an on-call external resource, is hired by hospitals to take care of patients who are at risk and need constant supervision. One of my research is to analyze the combined data of sitter usage and anonymous patient information and predict the potential sitter requirements (i.e., expertise and amount) that hospital may need later. The proposed method uses both of regular expressions and simple vector model and the predicted results can help managerial level to fine tune staff proportion to better respond patient needs.

Data is first collected from both of sitter administration and hospital patient tracking systems. The data from different systems are merged logically according to the patient's medical record numbers and hospital site. Once merged, data columns that contain any patient specific or any information that can lead to find anyone will be removed.

Depending on the software and database design, data quality issues may exist. When two data from different sources are combined, the issue can be amplified. Data collected from different systems is often dirty as it may be incomplete (e.g., have missing attribute values); it may have noises (e.g., have out of range values or exaggerated values that even do not make sense); it may have inconsistent data (e.g., free text user inputs); it may have duplicated records. Dirty data leads to poor data quality and may lead to ineffective data mining results (Han & Kamber, 2006). Data cleanup is then done to identify possible erratic entries and eliminate non-useful data. Some data is discretized into uniform ranges to facilitate data analysis.

Data normalization is important for clustering data since clustering classifies data into different group by observing the difference among data. However, some attributes may have greater value ranges than the others. For

example, the value range of the attribute length-of-stay can be 1 to several hundreds but attribute age-of-patient may only have range like 1 to 120. If data normalization is not applied before the data analysis takes off, attributes with greater value ranges may dominate the clustering results and the others would be "ignored".

The next step is data reduction that tries to reduce the data volume; for instance, date-of-birth attribute is retrieved from patient record. Despite a year has 365 days and patients could be born in any of day even in the same year. It may not have strong connection between a patient's health issue and a date in a year but may have connection between his or her age. Therefore, the date-of-birth attribute and its values are replaced with age attribute and its correspondent values. Furthermore, the age values can be grouped together based on preset criteria. Instead of having potentially 120 different ages, only dozen of age groups are being used.

A novel method of "predicting" sitter requirement such as the sitter's expertise is proposed. The method predicts sitter case attribute value with the techniques of sequence matching, regular expression and vector space model. I use some examples to briefly explain the method. For more details of the proposed method, please refer to (Lo & Chang, 2012).

Each sitter case may involve and be enriched the following information:

- Mission and hospital site where the patient is hospitalized
- Shift requiring the sitter service
- The patient's gender
- Type of admission
- The patient's Marital status
- Length of stay this information can be known from hospital patient tracking system
- Discharge location (e.g., home) this information can be known from hospital patient tracking system

The proposed method considers all sitter cases as sequential records based on their dates and shift stamps. The method needs to discover the similarity in-between two sequences.

Sequence can be generated for an attribute based on filtering criteria. The chosen attribute is considered as the "seed". Assuming the administrative staff chooses "sitter reason" as seed and makes "site=MGH, shift=Day, gender=Male, admission type=Clinic, and marital status=Single" as filtering criteria to predict the attribute value "length of stay", the method generates a reference sequence of sitter reasons based on the searching criteria. To facilitate the representation of the sequence element and make regular expressions applicable later, a single alphabet index is being used to represent each attribute value instead of a complete word. For instance, a sequence EJAJO stands for the reasons of a sequential sitter cases based on

the filtering criteria is: AwayWithoutLeave (E), Disorientation (J), Agitation (A), Disorientation (J), and Suicidal (O).

The method makes use of the sequence similarity to discover relationship between pre case and post case attribute values. In other words, it is assumed that the symbolic sequence of an attribute may contain hints to reveal other attribute values. For example, a series of sitter reasons can be used as a predictor to predict how long patients are going to stay in the hospital – length of stays.

Considering the example mentioned earlier above, the method then generates other sequences by further taking the values of attribute "length of stays" that the administrative staff intends to predict into the filtering criteria. Assuming attribute "length of stays" has three values, the generated sequences for the seed attribute "sitter reason" are JJJ, JJJJJJJJAAAAAALAAAAA, and AAAA.

Word matching technique is adopted to determine whether sequences are similar. A word is a series of elements in a sequence that is repeated. Regular expression based approach has been developed to find out possible sub-strings in different lengths. The distinct sub-strings found are treated as words.

Taking the generated sequence JJJJJJJJAAAAAAAAAAAA as example, the method can find all possible words with regular expressions' help:

 $\rightarrow (JJJJJJ)(JAAAA)(JAAAA)(AA)(L)(AAAAA)$

 \rightarrow (J+)(JAAAA)+(A+)L(A+)

Four distinct words are found. All of them are repeated at least once. Every time when a word is discovered, it is added into a dictionary.

Via the calculation of the frequencies that every words appeared in sequences and the use of vector space to present the frequencies, Cosine similarity (Zhao & Karypis, 2012) can be adopted to measure the similarities between the reference sequence and the other generated sequences. Cosine similarity has been widely used in clinical analysis to compare sequences generated by data collection tools with timestamps (Augustyniak, 2007; Bratsas, Hatzizisis, Bamidis, Quaresma, & Maglaveras, 2005; Chen, Hong, Huang, & Lee, 2008). It has also been proven to be a robust metric for scoring the similarity between two strings, and it is increasingly being used in text mining related queries (Subhashini & Kumar, 2010).

The method counts the number of occurrence of the words in each sequence. With all words' occurrence numbers, the generated sequences are represented by vectors of words. For instance, there are eight words found from the sequences. A vector of eight elements is created to represent each sequence with the numbers of word occurrences, as known as term frequency.

Since each vector contains exactly same elements, comparisons between two vectors can be done and which sequence is most similar to the reference sequence can be identified. The cosine of the angle between two vectors is a measure of how similar the two sequences are. Cosine of an angle can range from 0 to 1, where 0 means two vectors are not similar and 1 means two vectors are identical. The method identifies the sequence with the highest cosine similarity with respect to the reference sequence. The filtering value of the selected sequence is then the predicted result for the query. For cosine similarity with the 0.9818717 reference sequence sequences. Since the sequence came out while applying the second value (e.g., assuming B) that attribute "length of stays" has to the filtering criteria, the second value of the length of stays is the predicted value - next Single Male patient comes to Clinic at MGH in the Day shift may stay at the hospital for B days.

The research evaluates the proposed method with the data of all sitter usages within a hospital network that consists of five hospitals (4 adult sites and 1 child & adolescent site), for the entire years of 2008, 2009 and 2010. To evaluate the accuracy of the results predicted, the results are compared to the existing records in the databased. In general speaking, results are quite promising with fair accuracies.

4. CONCLUSION

In this chapter, I briefly talk three of my data analytics and more details can be further found and read from published conference and journal papers. The algorithm that identifies the difficulty of online quizzes and their items based on the frequent patterns extracted from students' behaviours of answering the quizzes and questions. Moreover, the found patterns may also show the course content designers (e.g., teachers or academic experts) that students may have difficulty in understanding specific part of the content or activities while learning due to the patterns are frequently happened and have very strong supports in the database.

The proposed algorithm can be further adopted by researchers and developers to design mechanism of finding users' common patterns while using any systems and help the system designers get clear idea of the possible user interface, workflow, and functions enhancements that their systems may be done. For instance, the intention of users can be figured out and its can be mapped to a pattern. If the pattern is a complicated one, then probably either the workflow or the user interface is not friendly for the users.

The mechanism adopted by Next-Stope Recommender mobile app can make suggestions of what its user should visit for next according to the similarity between his or her wandering behaviours and other users'. Although the mechanism was designed for providing personalized locationbased service running on mobile devices, it can also be used in the analysis of users' web browsing behaviours and provide users recommended webpages and information they may interest in the website, such as news recommendation service based on the results of matching the news read pattern of a visitor to others'. Students may also benefit from similar application such like external reading material recommendation service.

The regular expressions based prediction method help hospital administrative staffs predict the value of specific attribute that a forthcoming sitter case may have based on the chosen seed attribute and filtering criteria. The proposed method needs neither to know the meanings of attributes nor to do complicate calculations. It simply generates string based sequences, finds the words in the sequences, and measures the Cosine similarity between a testing sequence and the reference sequence. The method can further be applied to do prediction for the dataset from any disciplines and areas, as long as the dataset is sequential and the attributes chose as seed and target attributes are categorical or can be transformed to categorical attributes.

REFERENCES

- Alberta Education (2012). Bring Your Own Device: A Guide for Schools, Technology in
Schools. Retrieved June 3, 2015 from
http://www.k12blueprint.ca/sites/default/files/byod%20guide%20revised%202012-09-
05.pdf
- Alphonso, C. (2013). Canada's fall in math-education ranking sets off alarm bells. *The Globe* and Mail. Retrieved June 3, 2015, from http://www.theglobeandmail.com/news/national/education/canadas-fall-in-matheducation-ranking-sets-off-red-flags/article15730663/
- Augustyniak, P. (2007). Optimal Coding of Vectorcardiographic Sequences Using Spatial Prediction. *Journal of IEEE Transactions of Information Technology in Biomedicine*, 11(3), 305-311.
- Bhatia, R., Graystone, A., Davies, R., McClinton, S., Morin, J., and Davies, R. (2010). Extracting information for generating a diabetes report card from free text in physicians notes. In the Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, (NAACL HLT 2010), Los Angeles, USA, 8-14.

- Boudin, F., Nie, J., and Dawes, M. (2010). Positional Language Models for Clinical Information Retrieval. In the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Massachusetts, USA, 108-115.
- Bratsas, C., Hatzizisis, I., Bamidis, P., Quaresma, P., and Maglaveras, N. (2005). Similarity Estimation among OWL Descriptions of Computational Cardiology Problems in a Knowledge Base. *Journal of IEEE Computers in Cardiology*, 32(5), 243-246.
- Brochu, P., Deussing, M. A., Houme, K., & Chuy, M. (2013). Canada's students slipping in math and science. *OECD finds*, CBC news Canada. Retrieved June 3, 2015, from http://www.cbc.ca/news/canada/canada-s-students-slipping-in-math-and-science-oecdfinds-1.2448748
- Chapman, W.-W., Bridewell, W., Hanbury, P., Cooper, G.-F., and Buchanan, B.-G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*. 34(5), 301-310.
- Chapman, W., Chu, D., and Dowling, J. (2007). An Algorithm for Identifying Contextual Features from Clinical Text. In *the Proceedings of Biological, Translational, and Clinical Language Processing*, Prague, Czech Republic, 81–88.
- Chen, C.-M., Hong, C.-M., Huang, C.-M., and Lee, T.-H. (2008). Web-based Remote Human Pulse Monitoring System with Intelligent Data Analysis for Home Healthcare. In *the Proceedings of Cybernetics and Intelligent Systems*, (CIS 2008), 636-641.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Journal of ACM communications*, 39(11), 27-34.
- Grishman, R. (1997). Information Extraction: Techniques and Challenges. International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology (pp. 10-27), Rome, Italy.
- Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.
- Jia, Z., Li, H., Dong, L. and Long, D. (2011). Temporal Expression Recognition and Temporal Relationship Extraction from Chinese Narrative Medical Records. In the Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering, Wuhan, China, 1-4.
- Joyce, A. (2014). *Canadian Students are Massively Mobile*. Retrieved June 3, 2015, from https://www.londondrugs.com/on/demandware.static/Sites-LondonDrugs-Site/Sites-LondonDrugs-Library/default/v1427703642968/pdf/news/08-14-14_BTS_Cell.pdf
- Jurafsky, D., and Martin J.-H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Upper Saddle River, NJ: Prentice Hall.
- Konrad, R., and Lawley, M. (2009). Input modeling for hospital simulation models using electronic messages. In *the Proceedings of the Winter Simulation Conference*, Maryland, USA, 134-147.
- Li, B., Kuo, R., Chang, M., & Garn, K. (2015). Reward Points Calculation based on Sequential Pattern Analysis in an Educational Mobile App. In *the Proceedings of 21st International Conference on Distributed Multimedia Systems (DMS 2015)*, Vancouver, Canada. August 31-September 2, 2015, 186-190.
- Liu, D., & Chang, M. (2011). Next-Stop Recommendation to Travelers according to Their Sequential Wandering Behaviours. *Journal of Internet Technology*, 12(1), 171-179.
- Lo, S. H. K. and Chang, M. (2012). An Innovative Way for Mining Clinical and Administrative Healthcare Data. In *the Proceedings of Active Media Technology (AMT* 2012), Macau, December 4-7, 2012, 528-533.

- Mutalik, P.-G., Deshpande, A., and Nadkarni, P.-M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents. *Journal of the American Medical Informatics Association*, 8(6), 598-609.
- Ripley, B., Liu, D., Chang, M., & Kinshuk. (2013). Next Stop Recommender. In the Proceedings of 2013 International Joint Conference on Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA 2013), Aizuwakamatsu, Japan, November 2-4, 2013, 120-125.
- Seifert, J. (2004). *Data Mining: An overview*. Retrieved December 7, 2015, from http://www.fas.org/irp/crs/RL31798.pdf
- Subhashini, R., and Kumar, V. J. S. (2010). Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval. In *the Proceedings of the 1stInternational Conference on Integrated Intelligent Computing*, Bangalore, India, 27-31.
- Zhao, Y., and Karypis, G. (2012). Evaluation of hierarchical clustering algorithms for document datasets. In the Proceedings of the eleventh international conference on Information and knowledge management, (CIKM 2002), 515-524.
- Zhu, W., Fu, L., Xu, L., and Zhang, B. (2011). A TCM Diagnosis System Based on Textbook Information Extraction. In the Proceedings of the 4th IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing, Dalian, China, 483-487.

ACKNOWLEDGEMENTS

The authors wish to thank the support of Athabasca University and the Mission Critical Research funding. The authors also acknowledge the support of NSERC, iCORE, Xerox, and the research related gift funding by Mr. A. Markin.