2015

Proceedings of Science and Technology Innovations



Maiga Chang and Farook Al-Shamali

2015

Proceedings of Science and Technology Innovations



Maiga Chang and Farook Al-Shamali

Editors

Maiga Chang
School of Computing and Information Systems, Athabasca University
1200, 10011-109 Street, Edmonton, AB T5J-3S8, Canada
E-mail: maigac@athabascau.ca

Farook Al-Shamali
Centre for Science, Athabasca University
1200, 10011-109 Street, Edmonton, AB T5J-3S8, Canada
E-mail: farooka@athabascau.ca

Cover and Back page designer
Ying-Sheng Hu (supervisor - Prof. Wei-Peng Lien)
Department of Visual Design, Shu-Te University
59, Hengshan Rd., Yanchao District, Kaohsiung City, 824, Taiwan

ISBN 978-1-987973-00-6 (print)
ISBN 978-1-987973-01-3 (online)

This work is subject to **Creative Commons** license.

Publisher:
Faculty of Science and Technology,
Athabasca University,
1 University Drive,
Athabasca, AB T9S 3A3
Canada

Foreword

It is my great honour to introduce the first compilation of research accomplishments of Athabasca University's Faculty of Science and Technology students and faculty. In this new book, titled "2015 Proceedings of Science and Technology Innovations" you will explore research areas in ecology, computer science and data analytics, biology and science education. The work is innovative, inspiring and moreover, exhibits the close relationships we have in mentoring our undergraduate and graduate students within the Faculty of Science and Technology.

The faculty researchers in this edition are accomplished and talented, and are recognized both nationally and internationally. They are committed to making a difference in this world, and are resolute in their desire to lead by example for their students and in ensuring that they nurture the next generation of future scientists, researchers and decision-makers of our country and the world. This talent does not come easily, but when I observe the ways in which these individuals engage their students to consider questions from different perspectives and to apply state-of-the-art techniques, I am awed at their ability to guide and mentor their students, and to teach them how to use the power of curiosity in research.

As you read the ten chapters in this edition, the one thing that I would like to point out, is that these papers were written by our students with the support and guidance of our researchers. This process of encouraging student engagement fosters critical thinking, integrity, perseverance, resiliency and inquiry crafted around one or more questions around a certain topic. It allows our students to begin a journey of becoming our next leaders and role-models in science and computer science research and education.

I congratulate our researchers for this outstanding accomplishment and their leadership in research and teaching.

Lisa Carter

Dean, Faculty of Science and Technology
Athabasca University

Table of Content

1. **Identification of Low Nutrient Response Genes in the Bacterium *Pseudomonas aeruginosa* with Hierarchical Clustering** 1
   Bertrand Sodjahin, Shauna Reckseidler-Zenteno, Shawn Lewenza, Vive S. Kumar, and Junye Wang

2. **The Anomalous Behaviour of Water: A Simple Experiment** 17
   Farook Al-Shamali

3. **Integrated Modeling of the Athabasca River Basin using SWAT** ... 27
   Getnet D. Betrie, Baoqing Deng, and Junye Wang

4. **Improving Learning based on the Identification of Working Memory Capacity, Adaptive Context Systems, Collaborative Learning and Learning Analytics** 39
   Richard A.W. Tortorella, Darin Hobbs, Jeff Kurcz, Jason Bernard, Silvia Baldiris, Ting-Wen Chang, and Sabine Graf

5. **Enhancing Mathematical Problem-Solving Experiences through Learning Analytics** 57
   Rébecca Guillot, David Boulanger, Jérémie Seanosky, Vivekanandan Kumar, and Kinshuk

6. **Data Analytics for Education and Healthcare** 75
   Maiga Chang

7. Enhancing Music Prowess through Analytics..... 93
Claudia Guillot, Rébecca Guillot, Vivekanandan Kumar, and Kinshuk

8. Semi-Supervised Product Specifications Extraction from the Web
..... 105
George Krys and Ebrahim Bagheri

9. User Modeling for Course Planning and Scheduling 121
Fuhua Lin, M. Ali Akber Dewan, and Alex Newcomb

10. Multi-Agent Well Scheduling: A Prototype Implementation Using CNP and JADE..... 137
Jivko Hristov, Graham Lange, Fuhua Lin, M. Ali Akber Dewan, Xiaokun Zhang, and Saadat Khan

Author Index..... 155

Chapter 1

IDENTIFICATION OF LOW NUTRIENT RESPONSE GENES IN THE BACTERIUM *PSEUDOMONAS AERUGINOSA* WITH HIERARCHICAL CLUSTERING

Bertrand Sodjahin[1], Shauna Reckseidler-Zenteno[1,2], Shawn Lewenza[1,2], Vive S. Kumar[1], Junye Wang[1]
[1] *Faculty of Science and Technology, Athabasca University, Canada*
[2] *Department of Microbiology, Immunology & Infectious Diseases, University of Calgary, Canada*

Abstract: *Pseudomonas aeruginosa* is a bacterial organism known for its ubiquity in the ecosystem and for its ability to resist antibiotics. It can survive at length in any environment it reaches, in particular hospital surfaces and is deemed to cause various diseases, in humans, animals, and plants. It is a common cause of nosocomial, hospital-acquired infections. It has been shown that this organism can be isolated from water in a number of intensive care units. The hypothesis is that *P. aeruginosa* is capable of long-term survival in water due to the presence of particular genes which encode for proteins that facilitate persistence. The objective of our research is then to identify genes involved in the survival of *P. aeruginosa* in water by looking at genes responsive to a low nutrient environment. We conducted on a gene expression data a hierarchical clustering analysis in Weka, which is a collection of machine learning algorithms for data mining tasks. The results appear to be interesting, yielding a list of 91 distinct genes accounting for approximately 8% of the genome and identified as potentially responsible for the survival in water of the bacterium.

Key words: Machine Learning, Hierarchical Clustering, genes expression, *Pseudomonas aeruginosa*, Weka.

1. INTRODUCTION

Pseudomonas aeruginosa is a gram-negative bacterium that is ubiquitous in the environment and is known for its ability to inhabit a number of environments, causing disease in plants, animals, and humans (Jørgensen et al., 1999). This diverse organism is also a common cause of hospital-acquired infections, mostly causing skin infections in burn patients, infections of indwelling devices such as catheters, and fatal lung infections in patients with cystic fibrosis (Driscoll, Brody, & Kollef, 2007). Studies have shown that *P. aeruginosa* may survive for months on hospital surfaces (Kramer, Schwebke, & Kampf, 2006). Infection by this bacterium is very difficult to treat because of its resistance to a number of antibiotics (Driscoll et al., 2007). It utilizes certain mechanisms to resist the effects of antibiotics including efflux pumps, modification of the outer membrane to reduce permeability, and inactivation of drugs through the production of enzymes (Driscoll et al., 2007). One of the most effective ways of combating the effects of antibiotics is for the organism to exist as a biofilm (Driscoll et al., 2007; Harrison, Turner, & Ceri, 2005; Ryder, Byrd, & Wozniak, 2007) which is the result of a complex aggregation of microorganisms surrounded by a protective and adhesive matrix. These biofilms are dramatically more resistant, up to 1000 fold, to antibiotic treatment due to the protection provided by the surrounding matrix polymers (DNA, protein, polysaccharides), the slow growth rates of nutrient limited cells within a biofilm and the presence of multidrug tolerant persister cells (Harrison et al., 2005). We are interested in understanding how *P. aeruginosa* is able to survive in the environment, particularly in water. The ability of the organism to persist at length in water without any nutrients may be responsible for its introduction in hospital environments, leading to patients infections. Not only is *P. aeruginosa* an important opportunistic pathogen and causative agent of nosocomial infections, it can also be considered a model organism for the study of diverse bacterial mechanisms that contribute to bacterial persistence. One of the reasons for its ability to survive in a number of conditions may be due to the large genome it possesses (Stover et al., 2000; Wolfgang et al.,, 2003). The presence of a large number of genes, 50% more genes than *E. coli* permits diversity and adaptability by the organism. Our research goal is to identify these genes involved in the survival of *P. aeruginosa* through the genes' response to low nutrient water. The paper is organized as follows. Methodology and underlying literature in section 2 first describes the data and the analysis environment. Then it discusses the data pre-processing methods as well as the analysis configuration and

implementations in Weka. In Section 3, we present our results and discussions. Section 4 is consecrated to the conclusion of our current work and to some details concerning our future works.

2. METHODOLOGY AND UNDERLYING LITERATURE

2.1 Data description and analysis environment

Fundamental study interests in genetics and microbiology mostly concern functional genomics, genes sequencing, gene profiling or genes expression level for the identification of genes associated for instance with a certain phenotype manifested by an organism. Molla, Waddell, Page, & Shavlik, (2004) introduce genes as components of DNA which encodes for protein and define gene expression as the sequential steps of the transcription of the DNA, which it is part of, into RNA and the translation of this latter into associated protein. In other words, the expression level of a gene is measured with as proxy, the observation of protein fabrication rate in an organism which in response to its environment switches on or off its protein production. Though measuring the expression of an individual gene has been previously achieved, it wasn't until the advent of microarray technology that simultaneous expression measurements of thousands of organism's genes are made possible. Babu (2004) described microarray as a glass slides assembly in which DNA molecules are orderly fixed at particular places referred to as spots or features, each containing millions identical copies of DNA molecules corresponding uniquely to a gene. This microarray technology makes it easy to capture at once integral biological activities and therefore conducive to obtaining high-throughput data, useful for example in the inference of cells regulatory pathways. One of the predominant uses of microarray in gene expression is in the comparison of expression measure of a set of genes originally maintained under a certain condition, with the same set under different other conditions. This permits the study of the impact of these conditions on gene expression.

In our current research study, to identify the genes involved in the survival of *P. aeruginosa* without nutrients, an existing transposon library of *P. aeruginosa* mutants was utilized. This mini-Tn5-*luxCDABE* transposon mutant library of *P. aeruginosa* PAO1 is a collection of random transposon mutants, each containing a mutation in a different gene. This is the result of insertion of a mini-Tn5 transposon into the gene, which prevents effective transcription and eventually translation of the gene into a functional protein.

Each insertion of the mini-Tn5 transposon contains the *luxCDABE* operon, which results in light production as the gene is being transcribed. This allows for determination of gene expression under a variety of conditions. The *luxCDABE* operon is derived from the bacteria *Photorhabdus luminescens*, which is a luminescent marine bacterium (Winson et al., 1998). The mini-Tn5-*luxCDABE* library in PAO1 contains 9,000 mutants, 2,500 of which have been mapped and characterized. Of the 2,500 characterized mutants, 1,384 of these were determined to produce light. This collection of 1,384 mutants was screened for gene expression in water and the gene expression data (Lewenza, Kobryn, de la Fuente-Nunez, & Reckseidler-Zenteno, 2015) has therefore approximately 15,000 data points to be analyzed. It has overall 15 columns. The first of which is the well ID that in fact represents the array identification of wells in which the mutants have been inoculated. The second column is the gene, the third is the PA number, the fourth is the product name, the fifth is the original well ID before the transfer, and columns 6 to 15 (T_4 – T_{672}) represent the ten different time points of the gene expression which represents the ratio of the actual measurement (absolute value) at time at time T_i ($i>0$) by the value for the same gene at time zero (T_0). This ratio establishing procedure is known as normalization (from the absolute value to a relative value). Its use is justified by the fact that, accurately estimating the absolute expression level of certain genes is challenging (Molla et al., 2004). So normalization is a way of canceling systematic variations that are induced by various sources such as different amount of starting mRNA material in two examples (Babau, 2004). Because gene expression matrix may be made up of absolute value or relative value, in order to prevent erroneous analysis, one must always first identify the type of values (absolute or relative) contained in a gene expression matrix, before undertaking any processing step. At this stage of our work, we focus mostly on the identification of the genes responsible for the persistence of the bacterium in low nutrient water.

For this analysis, among software and platforms we've explored are SPSS, a software package used for statistical analysis; AMOS, a statistical software package for structural equation modeling (produced by SPSS); and R, a software environment for statistical computing and graphics. One that appears to be the most practical and suitable is Weka, for several reasons. Weka environment, endowed with friendly usability, is a collection of machine learning algorithms (which include classification, regression, clustering and association) used to mine data. These embedded algorithms can either be applied directly to a dataset or called from one own Java code. In addition to the algorithms, Weka contains data pre-processing and visualization tools. Though we are not using its development components at this stage, it is actually well-suited for devising machine learning schemes. These make Weka of first choice for our future development given its compatibility with

the two main stream Operation Systems of reference: Mac OS and Windows. The latter is the one we use here.

2.2 Pre-processing methods and literature

The data raw described above (Lewenza et al., 2015) is typically a microarray data that includes more information columns than we need. The first step of our work consists in trimming it down to only the columns of genes and those of the 10 time point gene expression measures (table 1). This brings our dataset to the form that Babu (2004) classified as gene expression matrix's relative measurement. Discussing the various representations of gene expression data, Babu (2004) actually discussed discretization of the time point measurement. Intrinsically, most physical measurements are continuous and discretizing them is often needed, not only based on applications requirements but also as a mean of noise cancellation. Whether gene expression matrix is in absolute or relative representation, it can be discretized.

In our present study case, the values of the gene expression measurement are all numerical (Real numbers). As in gene expression analysis studies in general, discretization is required here as well. We base it on the significance of expression level measure as threshold according to which we qualify a gene as expressed or repressed. We define a variable *Gene_Express_Val* as the gene expression measurement value. We discretize the values by representing every value equal or greater than 2 (2-fold) as "YES" meaning expressed or up-regulated, and "NO" for those that are below i.e., down-regulated. Though this pre-processing can be partially done in Weka which offers such tools, we used excel. One main reason for this is that our original data is in Excel format and it is easier to remove in Excel columns that are not relevant to our work at this time. Those that we kept include genes column and all the 10 time point gene expression measurement columns which we discretize in excel into binary nominal attributes (YES, NO) through Algorithm 1.

Algorithm 1: Our Pseudo code for the gene expression discretization

```
1. If (Gene_Express_Val >= 2) {
2. Gene_Express_Val := YES;
3. } Else if (Gene_Express_Val < 2){
4. Gene_Express_Val := NO;
5. } Else {
6. Do nothing; // this is for missing values with "?"
7. }
```

Table 1. Sample of the trimmed down microarray gene expression matrix.

Genes	**T4**	**T8**	…	**T672**
PA5398	1.926365	1.427299	…	0.030316
PA5400	2.138769	1.51678	…	0.048796
…	…	…	…	…
?	1.066667	1.390476	…	0.819048
…	…	…	…	
Tgt	1.06	0.64	…	0.73

Table 2. Sample of the pre-processed microarray gene expression matrix.

Genes	**T4**	**T8**	…	**T672**
PA5398	NO	NO	…	NO
PA5400	YES	NO	…	NO
…	…	…	…	…
?	NO	NO	…	NO
…	…	…	…	…
Tgt	NO	NO	…	NO

There are missing data in some cells of the original dataset. Though we could use k-nearest neighbour (kNN) method for data imputation as in Low et al., (2014, October), due to the minimalistic number of missing data, we just managed this by filling in within Excel the empty cells with "?" to denote unknown value more specifically non-identified gene.

Babu (2004) has highlighted some key terminologies that are very useful in referencing portions of a data in the gene expression matrix. The first is *gene expression profile* which corresponds to the cumulative expression levels for a gene across all the experimental conditions. The second, *sample expression profile* alludes to the cumulative expression levels for all the genes in a single experimental condition. *Vectors space* is another representation alternative of gene expression data. It is in fact a mathematical concept domain in which gene expression profiles and sample expression profiles are represented respectively as horizontal and vertical vectors. Such domain is very useful in applications involving matrix operations in certain data processing procedures such as the rotation of an image by an angle α given its data matrix represented as vectors space. This representation is actually what we have in table 2 and because our study

concerns with genes pattern identification, we concentrate on *gene expression profile*.

Because our data file is in Excel (sample illustrated in table 2), and Weka requires input format of *.ARFF* file, we first converted the Excel file into *.CSV* format and then use an online CSV to ARFF converter called **csv2arff** to format our file into *.ARFF*. The resulting file is then used as the main data input for our analysis.

2.3 Analysis configuration in Weka: methods selection and literature background

A rich review of algorithmic techniques for gene expression data analysis is conducted by Kerr, Ruskin, Crane, and www.wcci2016.orgDoolan, (2008). Interestingly enough, they noted that selecting a method that best fits an experimental dataset is not without challenge. In other words there is no panacea method for all data. So this selection process has to be carefully carried out to obtain a technique that yields optimized results.

Molla et al. (2004) in their work on using machine learning to interpret gene-expression microarray in biological applications distinguished supervised learning and unsupervised learning. In our dataset, the genes are not already categorized or labelled and our research goal is to isolate group of genes that are responsible for the persistence of *Pseudomonas aeruginosa* bacterium. Therefore unsupervised learning is suitable for our analysis. Under the unsupervised learning Molla et al. (2004) discussed two main groups of learning algorithms: Clustering and Bayesian Networks. Both of these groups of learning algorithms are of key interest to our research work. First, clustering methods consist in grouping or clustering examples provided through a dataset, from which it learns by evaluating the similarity of their feature values, notably gene-expression values in our case here. According to Molla et al. (2004), the flexibility and intuitiveness of clustering make it widely adopted by biologist researchers and is well used in the domain of bioinformatics. For instance, Do and Choi (2008) surveyed the basic principles of clustering DNA microarray from various clustering algorithms. Babu (2004) broadly divided clustering methods in two majors groups: Hierarchical and non-hierarchical, though they are much more complex in their categorization as shown in table 3, which is excerpted from Han, Kamber, and Pei (2011a). Bayesian Network, the second type of learning algorithms, is to be considered in the next stage of our research and we will briefly discuss it in our future work section.

As far as method and process that we proposed here are concerned, after the pre-processing that resulted in the data format observed in table 2, we

need to proceed to its clustering. One of the primordial requirements for clustering analysis is the determination of the number of clusters which is generally challenging and requires domain knowledge. Based on our dataset, the goal of our research (isolate genes responsible for *P. aeruginosa* survival under low nutrient environment) and our background knowledge, we intuitively posit we need two clusters. This implies that genes that are not part of this group of interest will form a second group. Therefore through our analysis, the genes are to be partitioned in two groups of respectively similar gene expression pattern. Among the various clustering methods as shown in figure 1 that is excerpted from Chaudhari and Parikh, (2012), hierarchical appears to be the most suitable at this stage of our study work. First because by visually and "humanly" looking at the data, we forebode a certain hierarchy and interaction as we observe that genes which were initially down-regulated become up-regulated at a later time and vice versa, across time. Of previous researchers that have abounded in the same direction of hierarchical clustering, are Eisen, Spellman, Brown, and Botstein (1998) who used this technique to repeatedly pair two most similar examples, for the grouping of genes based on similarity in their expression pattern.

Indeed, as part of the gene expression data analysis is the distance measure which is the quantification of similarity between the sample objects under consideration, here genes. For the computation of similarity or dissimilarity, under hierarchical clustering in Weka, we have options between Euclidian, Manhattan, Minkowski, and Chebyshev distances. Though Euclidian distance is known to be of the most popular usage in clustering algorithm, we choose Chebyshev distance. It is a generalization of Minkowski distance which is itself a generalized form of Euclidian and Manhattan distances respectively (Han et al., 2011b) and it attributes equal distance to all its height neighbours as per its grid representation (Figure 1). So we choose the most generalized form, **Chebyshev distance**. Integral part of this clustering algorithm are also the distance approach considerations: *single linkage* (smallest distance), *complete linkage* (longest distance), *average linkage* (average distance) and *centroid linkage* (center distance). These are to be selected based on the clustering objectives and the domain. In fact computing Chebyshev distance is finding attribute f=f_{max} (among the P attributes in the dataset) for which the distance between two objects (here genes) is maximum. This definition makes our distance approach selection to be the **Complete Linkage**. Given two genes *Gi* and *Gj*, and based on Chebyshev distance formula from Han et al. (2011b), we write the equation (1) below that computes the Chebyshev distance between these genes.

$$d(Gi, Gj) = max_f^P \left| f_{Gi} - f_{Gj} \right| \quad (1)$$

The hierarchical methods, as observed in figure 2, is subdivided into agglomerative (bottom up) and divisive (top down) clustering. While the first proceeds through an iterative agglomeration of individual genes till all genes form a single cluster, the second uses an iterative division till each gene forms a group of its own. In Weka, as noted in the synopsis, the class *weka.clusters.HierarchicalCluster* is agglomeration based and is therefore what we use. Its general algorithm is presented in Algorithm 2.

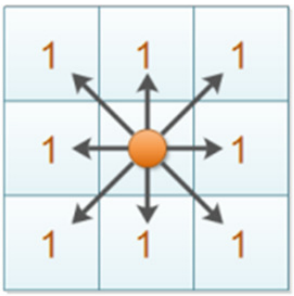


Figure 1. Chebyshev distance grid

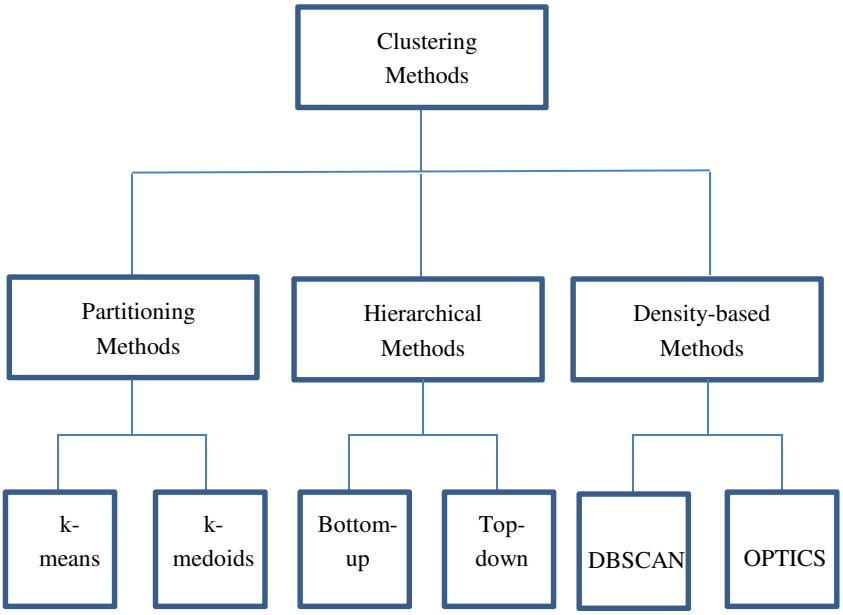


Figure 2. Graphical Overview of clustering methods

Algorithm 2: Basic agglomeration clustering algorithm

1. Compute the distance matrix, here **Chebyshev**
2. **Repeat**
3. Merge the closest two clusters
4. Update the distance matrix to reflect distance between new clusters
5. Until only one cluster remains

Table 3. Description of clustering methods.

Methods	**General Characteristics**
Partitioning Methods (e.g., K-Mean)	- Find mutually exclusive clusters of spherical shape - Distance-based - May use mean or medoid (etc.) to represent cluster center - Effective for small- to medium-size data sets
Hierarchical Methods	- Clustering is a hierarchical decomposition (i.e., multiple levels) - Cannot correct erroneous merges or splits - May incorporate other techniques like microclustering or consider object "linkages"
Density-based Methods	- Can find arbitrarily shaped clusters - Clusters are dense regions of objects in space that are separated by low-density regions - Cluster density: Each point must have a minimum number of points within its "neighborhood" - May filter out outliers
Grid-based Methods	- Use a multiresolution grid data structure - Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

3. RESULTS AND DISCUSSIONS

In our work while conducting various analysis methods experiments in Weka on the same dataset, we realized as discussed in Do and Choi, (2008) that with different clustering algorithms, similarity metrics and number of clusters, results vary substantially. Our approach had then been to first have a theoretical review of the methods and with our background knowledge to

select the suitable parameters and practically test them to see the one that provides the best results and this is what we did in section 2.3.

We have in figure 3 the results of the clustering as well as the parameters used. For instance it can be observed that **Hierarchical clustering** is used with **Complete Linkage** distance calculation. In fact based on our experiment, Complete Linkage provides the best clustering compared to the other distance approaches discussed in section 2.3 which were unable generate any result. We see in the same figure 3 that the percentage repartition shows 92% in cluster 0 and 8% in cluster 1. In other words, this means that 92% on one hand have similar expression profile while 8% on the other hand share same expression pattern. Mesquita, Soares-Castro, & Santos, (2013) noted that a comparative genomic analysis of *Pseudomonas aeruginosa* revealed it could be considered as a mosaic of two components and Kung, Ozer, & Hauser, (2010) evaluated the core at approximately 90% of the total genome and by implication the accessory nears 10%. Therefore our result of 92% for cluster 0 and 8% for cluster 1 not only allows us to draw the conclusion that the smallest percentage of our results contains the group of genes that have similar expression pattern responsible for the survival in low nutrient environment. For this group, our percentage obtained is approximately equal to those of Kung et al., (2010) in which it was further pointed that "the accessory genome may encode gene products that contribute to the niche-based adaptation of the bacterium, such as increase in host range, survival in new environment and utilization of new nutrients". This comes in a good alignment with the outcome of our results which makes us conclude that the 8% of the overall genes contains the genes we purposed to identify and which are potentially responsible for the persistence of the of *Pseudomonas aeruginosa* bacterium. And we should further zone in into this restricted group using additional biology information genes of this group that might be outliers

We visualized the clusters and obtained figure 3 showing red cluster elements as belonging to cluster1 and blue ones as belonging to cluster 0. We notice that the line is very fine between the two clusters and not easily separable when we observe the graphic. So we actually saved the result which generated another .ARFF file. This file comes out as a modified version of our input file that we loaded in Weka. It now includes an additional attribute which is named Cluster (see table 4) and is of nominal type (cluster0, cluster1). For each data tuple in this file, cluster 0 or cluster 1 has been added to classify each gene as result of our analysis. We then converted this file into CSV and opened it in Excel where we applied A-Z sorting by cluster column in order to group tuples with cluster 0 in the first rows and cluster 1 in the last rows. This allows us to copy at once the genes

labelled cluster 1 as result of our analysis. We also removed duplicates as well as non-gene rows. This as result allows us to isolate list of 91 genes.

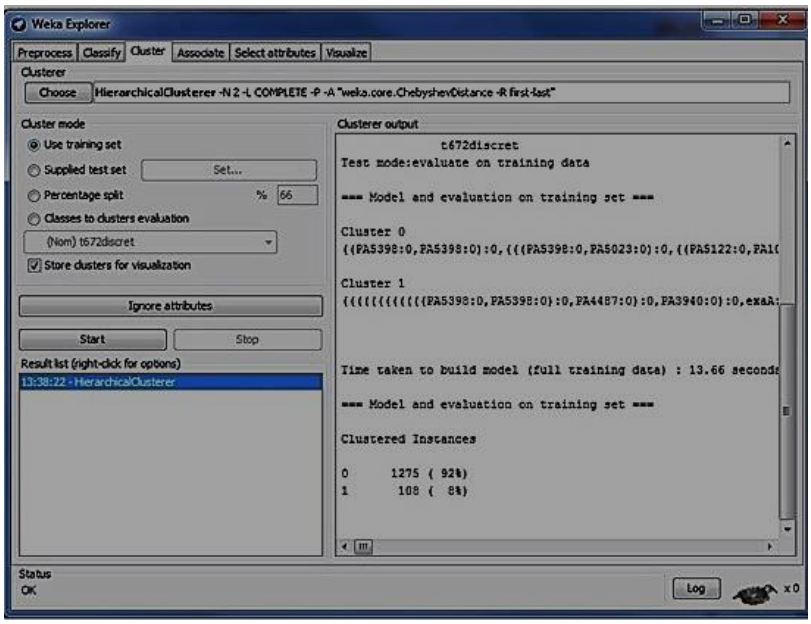


Figure 3. Parameters used and clustering results

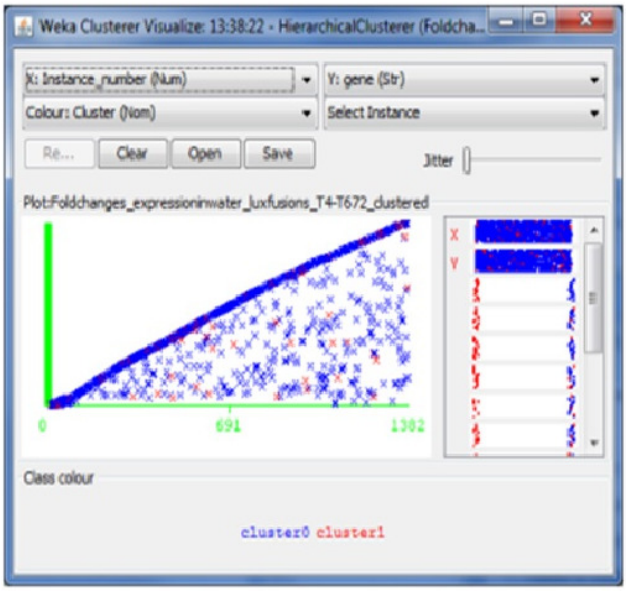


Figure 4. Clusters visualization

Table 4. Sample of the analysis result matrix with cluster labels.

Genes	**T4**	**T8**	…	**T672**	**Cluster**
PA5398	NO	NO	…	NO	cluster1
PA5400	YES	NO	…	NO	Cluster1
…	…	…	…	…	…
?	NO	NO	…	NO	…
…	…	…	…	…	…
Tgt	NO	NO	…	NO	Cluster0

Not only are other types of clustering found in the literature such as such as K-means (Yeung, Fraley, Murua, Raftery, & Ruzzo, 2001) that have proven successful though not suitable in all cases, there are also non-clustering methods such anti-clustering filtering (Raza, & Mishra, 2012) and time series based methods such as dynamic Bayesian Network (Low et al., 2014) that is the next target in our future work.

4. CONCLUSION AND FUTURE WORK

4.1 Conclusion

In this work whose objective is to identify gene or group of genes responsible for the survival of *Pseudomonas aeruginosa* bacterium in low nutrients water. We used Machine Learning, especially hierarchical clustering with Chebyshev and complete linkage as similarity distance calculation metric in Weka environment. We identified a group of 91 distinct genes, representing about 8% of the overall genome of the *Pseudomonas aeruginosa* in the dataset studied. As discussed in our previous section (section 3), these interesting results of the 8% i.e., 91 genes isolated as potential source of survival in low nutrient water would be verified in our future work with more advance methods and infer the underlying survival mechanism. Because additional work are still to be carried out, we do not published here the full list of the 91 genes identified.

4.2 Future work

The above proposed method is the first stage of works yet to be pursued. We will use Bayesian network as a second level of analysis. Bayesian network because like clustering it is an unsupervised learning algorithm that is suitable to our dataset. Its probabilistic approach of processing is of interest to our analysis given the uncertainty involved in the genes interaction. Also according to Molla et al., (2014) the application of learning Bayes' nets to gene expression microarray data has drawn much attention due to the insight it provides pertaining to the interaction networks within cells that regulates gene expression. Within the Bayesian network methods, we find that Dynamic Bayesian network would be a best fit for our research. First because it is time series based and our dataset is a record of gene expression measurement at different time points. So we would use our data to infer a temporal direction for the interaction among the genes and would therefore highlight causal relations. Additionally we will investigate possible additional data source which would provide background knowledge for our analysis as it is known that prior or background knowledge is quite useful in improving learning. We will not only compare the performance of Clustering and Bayesian network for this dataset (as well as other existing methods) but we will also evaluate which one is the best in learning and accurately identifying the genes responsible for the persistence. Another important task for our research is to further investigate (beyond the identification of genes responsible for the persistence of *Pseudomonas aeruginosa*) what new knowledge can be acquired from the data. Our approach here is to use the power of Induction Logic Programming (ILP). The rules of ILP are easily interpreted by human and this makes it popular and well accepted in domains other than computer science. King et al. (2009) is a good example of the power of ILP which they used to include experiment design to devise an autonomous scientist which discovered new knowledge about functional genomic of yeast.

REFERENCES

Babu, M. M. (2004). Introduction to microarray data analysis. Computational Genomics: Theory and Application, 225-249.

Chaudhari, B., & Parikh, M. (2012). A Comparative Study of clustering algorithms Using weka tools. *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 1*(2).

Driscoll, J. A., Brody, S. L., & Kollef, M. H. (2007). The epidemiology, pathogenesis and treatment of Pseudomonas aeruginosa infections. *Drugs*, 67(3), 351-368.

Do, J. H., & Choi, D. K. (2008). Clustering approaches to identifying gene expression patterns from DNA microarray data. *Molecules and cells*, (25), 279-88.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863-14868.

Han, J., Kamber, M., & Pei, J. (2011). Cluster Analysis: Basic Concepts. *Data mining: concepts and techniques: concepts and techniques* (pp. 443-495). Elsevier

Han, J., Kamber, M., & Pei, J. (2011). Getting to Know Your Data. *Data mining: concepts and techniques: concepts and techniques* (pp. 39-82). Elsevier.

Harrison, J. J., Turner, R. J., & Ceri, H. (2005). Persister cells, the biofilm matrix and tolerance to metal cations in biofilm and planktonic Pseudomonas aeruginosa. *Environmental Microbiology*, 7(7), 981-994.

Jørgensen, F., Bally, M., Chapon-Herve, V., Michel, G., Lazdunski, A., Williams, P., & Stewart, G. S. A. B. (1999). RpoS-dependent stress tolerance in Pseudomonas aeruginosa. *Microbiology*, 145(4), 835-844.

Kerr, G., Ruskin, H. J., Crane, M., & Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in biology and medicine*, 38(3), 283-293.

Kramer, A., Schwebke, I., & Kampf, G. (2006). How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. BMC infectious diseases, 6(1), 130.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., ... & Clare, A. (2009). The automation of science. Science, 324(5923), 85-89.

Kung, V. L., Ozer, E. A., & Hauser, A. R. (2010). The accessory genome of Pseudomonas aeruginosa. *Microbiology and Molecular Biology Reviews*, 74(4), 621-641.

Lewenza S., Kobryn M., de la Fuente-Nunez C., & Reckseidler-Zenteno S. L., In preparation.

Low, S. T., Mohamad, M. S., Omatu, S., Chai, L. E., Deris, S., & Yoshioka, M. (2014, October). Inferring gene regulatory networks from perturbed gene expression data using a dynamic Bayesian network with a Markov Chain Monte Carlo algorithm. *In Granular Computing (GrC), 2014 IEEE International Conference on* (pp. 179-184). IEEE.

Mesquita, C. S., Soares-Castro, P., & Santos, P. M. (2013). Pseudomonas aeruginosa: phenotypic flexibility and antimicrobial resistance.

Molla, M., Waddell, M., Page, D., & Shavlik, J. (2004). Using machine learning to design and interpret gene-expression microarrays. *AI Magazine*, 25(1), 23.

Raza, K., & Mishra, A. (2012). A novel anticlustering filtering algorithm for the prediction of genes as a drug target. arXiv preprint arXiv:1211.2194.

Ryder, C., Byrd, M., & Wozniak, D. J. (2007). Role of polysaccharides in Pseudomonas aeruginosa biofilm development. *Current opinion in microbiology*, 10(6), 644-648.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., ... & Olson, M. V. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406(6799), 959-964.

Winson, M. K., Swift, S., Hill, P. J., Sims, C. M., Griesmayr, G., Bycroft, B. W., ... & Stewart, G. S. (1998). Engineering the luxCDABE genes from Photorhabdus luminescens to provide a bioluminescent reporter for constitutive and promoter probe plasmids and mini-Tn5 constructs. *FEMS microbiology letters*, 163(2), 193-202.

Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., Yang, Q., ... & Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of Pseudomonas aeruginosa. Proceedings of the National Academy of Sciences, 100(14), 8484-8489.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics, 17(10), 977-987.

Chapter 2

THE ANOMALOUS BEHAVIOUR OF WATER: A SIMPLE EXPERIMENT

Farook Al-Shamali
Faculty of Science and Technology, Athabasca University, Canada

Abstract: A perception exists among students, and many instructors, that highly quantitative physics experiments can only be done in supervised physics laboratories and using specialized and costly equipment. This restrictive view poses an obstacle for the development of correspondence/online courses. With good imagination and adequate research, high quality physics experiments can be designed and performed safely and independently by the students at home using common materials and low-cost devices.

With this in mind, a simple procedure is proposed at the introductory physics level that allows students to observe and perform a quantitative study of the anomalous expansion of water. The procedure involves monitoring the volume of water below and above the anomalous temperature including some interesting observations such as the super-cooled state of water. The temperature at which water reaches its maximum density is estimated in the experiment. For the setup, students use commercially available equipment and common household items, and perform quantitative measurements that are comparable to traditional physics labs.

Key words: Physics, Home-lab, Experiments, Anomalous temperature

1. INTRODUCTION

The temperature of a gas or a liquid, based on the kinetic theory, is a measure of the average translational kinetic energy of the molecules. Therefore, raising the temperature of a liquid (for example) should increase the average

separation between the molecules, leading to lower the density. The same principle applies to water, however, with a very interesting exception observed at low temperatures near the freezing point. At about 4°C water reaches its maximum density. Beyond this temperature water will always expand, whether it is being warmed up or cooled down.

This anomalous behaviour is an important characteristic of water, which played a significant role in sustaining marine life. Because of this unusual property, freezing of a body of water, due to changes in weather, begins at the surface. Since ice is also less dense than water, it floats forming an insulating layer that considerably slows thermal conduction from the warmer water below to the colder air above. As a result, life in relatively large fresh water lakes (for example) is protected from the harsh winter weather above and continues in liveable water temperature near 4°C (Soletta & Branca, 2005).

The observation and measurement of this anomalous behaviour near the freezing point is a very interesting and informative experiment in physics (and chemistry). However, incorporating a suitable experiment in an introductory lab course has some procedural and technical challenges. This is simply because the coefficient of volume expansion of water at low temperatures is very small. The water volume only changes by 0.013% in the temperature range 0°C – 4°C (De Paz, Pilo, & Puppo, 1984).

In our case, we had extra technical and financial restrictions, concerning the setup, caused by our intent to include these experiments in the home-lab components of distance education physics courses at Athabasca University (Al-Shamali & Connors, 2010). Therefore, the setup had to be inexpensive, transportable and doable. Also, it was important to ensure that the student can conduct the experiment at home independently and safely without compromising lab quality and acceptable learning outcomes. With these requirements in mind, we designed the experimental procedure discussed in this article.

2. DENSITY AND TEMPERATURE

Since its appearance in the 1960's, the lava lamp became a familiar fixture and a popular house decoration item (Leif, 2008). A close observation of this device provides a clear demonstration of the relationship between density and temperature of a liquid. The lamp contains two insoluble liquids with slightly different densities. Buoyant forces act to achieve a stable layering with the denser liquid at the bottom. This stability is disturbed by turning on a lamp that heats the liquid at the bottom, which interns expand, until its density becomes less than that of the liquid above and starts to rise. The temperature

of the rising liquid, however, decreases as it moves away from the heat source and, eventually, becomes the denser liquid again and starts descending.

Considering a column of a single liquid, the temperature (and therefore the density) is not expected to be uniform throughout the liquid. However, buoyancy always acts to rearrange density distribution in a vertical continuum (or gradient) starting with the smallest density at the top to the largest density at the bottom of the column. Since density usually decreases with increasing temperature, the stable state also means a continuous temperature gradient from the warmest liquid at the top to the coldest liquid at the bottom.

Water, in particular, has the same general behaviour for the majority of the possible temperature range down to 4°C. At this temperature, however, water reaches its maximum density and expands as it gets colder. Therefore, away from this temperature, a water column should remain statically stable (no convection) during a gradual decrease (or increase) of its overall temperature. However, when the temperature inside the column starts to cross the 4°C mark, the stable density distribution is disturbed generating convection currents. Even though, the change in water density in the temperature range (4°C – 0°C) is relatively very small, it causes a dramatic reversal of the whole temperature gradient in the water column in order to revert to the stable density distribution. As a result, the upper and lower ends of the water column undergo significant temperature changes while this reordering of colder and warmer layers of water takes place.

At this point, it should be emphasized that the main idea behind the experiment presented in this article is not new. Actually, relevant literature goes back as far as 1805 when Thomas Hope published a set of experiments to observe and study this interesting behaviour of water (Greenslade, 1985). In his experiments, Hope did not attempt to measure the water volume as a function of temperature. He, however, employed an indirect (but elegant) method that monitors the redistribution of the temperature gradient in a water column as a result of density changes around 4°C. Here, we present a modern and more practical experiment suitable not only for traditional undergraduate laboratories, but also for home-labs in distance education courses.ext section of your chapter begins here.

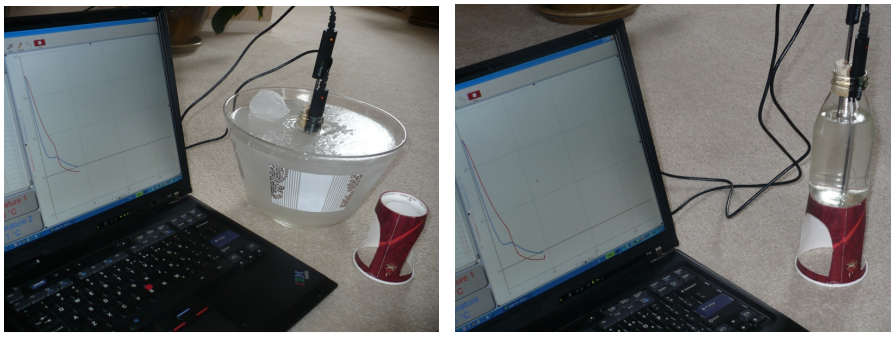
3. EXPERIMENT AND OBSERVATION

The main aspect of this experiment is the simplicity of its setup and procedure compared to the significant educational value and learning outcomes expected. Other than the commercially available Vernier temperature probe (Go!Temp), the experiment was conducted using common household items,

including a personal computer, a small size glass juice bottle, a deep kitchen bowl and some table salt.

The water column was formed by filling the empty juice bottle with tap water, avoiding the narrow section of the bottle neck. The cold reservoir was created by filling the kitchen bowl with salt-water solution and then cooling it down in a home freezer for few hours or overnight. Note that, depending on the freezer's temperature and salt concentration, the solution will become slushy or covered with a soft crust of ice. In this case it should be easy to make a hole in the crust that fits the bottle.

For temperature measurements, we used two probes that were inserted in the bottle at different heights (see Fig 1). The first probe, with its tip about 1 cm below the water surface, was sensitive to the temperature in the upper layer of the water column. The second probe, on the other hand, was pushed all the way down the water column such that its tip was just above (but not touching) the bottom of the bottle. The two probes were connected through USB ports to a personal computer. The "Logger Lite" data-collection software, installed on the computer, displayed the temperatures measured by the probes and allowed the simultaneous recording of the temperatures at both ends of the vertical water column. The software was set to auto-record the temperatures every 2 seconds.



(a) (b)

Figure 1: (a) Water bottle cooling down in a cold salt-water solution reservoir (b) The water bottle warming up by the surrounding air at room temperature.

Figure 2 displays two graphs corresponding to the temperatures measured (in one of the trials) by the two probes as a function of time. In this trial of the experiment, the auto data collection started with the initial water temperature inside the bottle at about 16°C. Note that the temperature at the bottom of the water column was initially less than that at the top. After few minutes, the bottle was inserted in the cold reservoir, with temperature near –10°C. As a result, the water temperature inside the bottle started to drop very quickly, as shown in the graph. It is interesting to note that while the top layer

of the water column remained warmer than the bottom layer, the temperature near the bottom cooled at a faster rate.

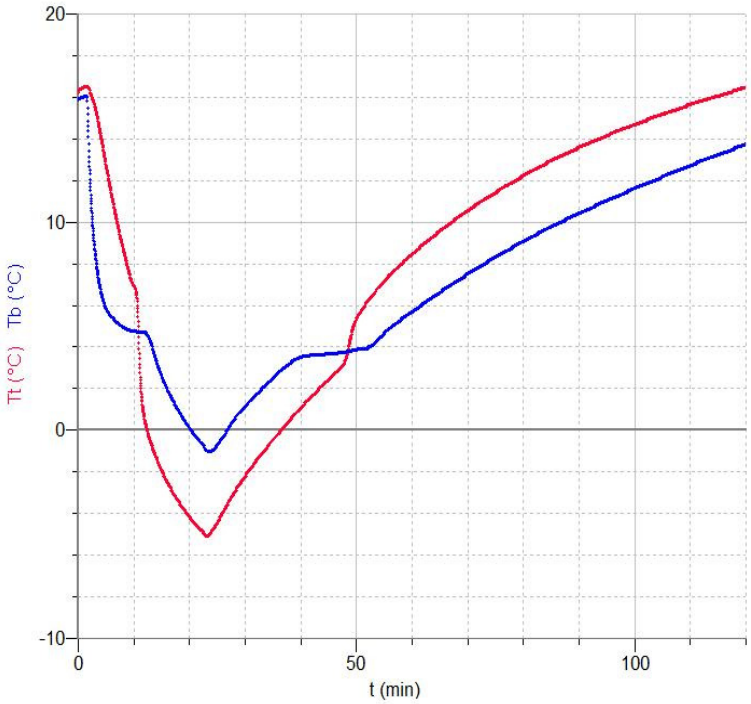


Figure 2: Temperatures of the top (red) and bottom (blue) layers of the water column inside the bottle.

When the temperature of the bottom layer reached about 6°C the cooling slowed down dramatically and nearly reached a plateau at about 4.5°C. This trend continued for more that 5 minutes before the temperature curve almost suddenly resumed cooling down. The temperature of the upper layer, however, dropped to about 7°C when it started to cool down much faster to a temperature of about 1°C. It is clear from the graphs that by this time the water near the top of the water column was colder than the water near the bottom, thus indicating a reversal in the temperature gradient of the water column. After that, the cooling continued (at a more normal rate) to subzero temperatures without freezing.

While the water column was in the super-cooled state, the bottle was carefully removed from the cold reservoir, dried out with a towel and allowed to warm up slowly by the surrounding air at room temperature. To minimize heat transfer irregularities, the bottle was placed on an inverted paper cup that had its bottom removed (see Fig 1b) thus allowing more uniform air-to-glass heat conduction from all sides of the bottle. As seen in the graph, the water

temperature started to increase, but at a slower rate than the cooling process. This is expected due to the differences in thermal conductivities between water and air in the two reservoirs.

In the warming process, the overall temperature trend (in Fig 2) was almost the opposite of what was observed during the cooling process. At about 4°C the warming rate of the bottom layer in the water column had a sharp decrease in the warming rate during which the temperature remained almost constant for about 13 minutes. The colder upper layer, on the other hand, continued to warm up at a normal rate until about 3.5°C. After that, the temperature increased more sharply to about 5°C and then resumed increasing at a more normal rate. At this time, it was obvious that the initial temperature gradient was restored with the warmer water near the top and the colder water at the bottom of the water column.

4. SUPER-COOLING

As mentioned in the introduction, the main objective of this experiment was to demonstrate the anomalous behaviour of water at low temperatures (above 0°C). However, it was also very interesting to note that the experiment provided an opportunity to demonstrate the phenomenon of super-cooling in which water cools down to sub-zero temperatures without freezing. Using clean drinkable water and avoiding shaking the bottle, we easily reached (see Fig 2) a temperature of –5°C inside the bottle. However, it should be noted that if the super-cooled state is not stable and if the bottle is given a good shake, instant freezing occurs and the temperature inside jumps to 0°C.

This phenomenon was also an opportunity to check the temperature probe's calibration and measurement uncertainty. Figure 3 shows the result of a test in which the two probes were inserted together inside the bottle such that their sensitive tips are at the same height in the water column. Ideally in this situation, the two probes should display the same temperatures. However the discrepancy in the measurement is within the probe's uncertainty (± 0.5°C) reported by the manufacturer. Of particular interest here is the sudden rise in the graphs, which coincides with the instant freezing of the super-cooled water in the bottle. The temperature, as measured by each probe, deviates from the expected zero value only by the equipment uncertainty.

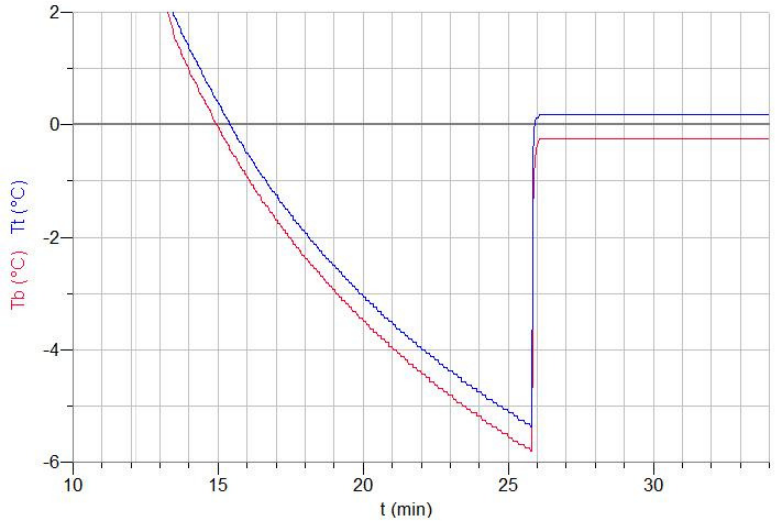


Figure 3: Cooling of the top (red) and bottom (blue) layers of the water column to the super-cooled state followed by instant freezing.

5. ANOMALOUS BEHAVIOUR

When the water column (or bottle) was placed in the cold reservoir, the temperature throughout the column started to drop very quickly, as seen in Fig 2. This is mainly due to thermal conduction, through the glass walls, from the warmer water inside the bottle to the colder water in the reservoir, according to the Second Law of Thermodynamics. From the temperature graphs, however, we notice that the bottom layer cooled faster than the upper layer in the water column. This can be explained by the argument that the relatively colder (and therefore denser) water formed just inside the bottle side walls convected down to the bottom and pushed the warmer water through the central column towards the upper layer. Therefore, while thermal conduction worked to reduce the overall temperature inside the bottle, convection currents accelerated the cooling process at the bottom of the column and slowed it down at the top. With this, we obviously disagree with the argument presented in (Gianino, 2007) to explain a similar observation referred to as the isothermal phase.

As the temperature of the bottom layer approached the anomalous temperature, the convection process mentioned above started to slow down, thus gradually reducing the cooling rate of the bottom layer. When the water near the bottom reached its maximum density the downward convection process at the side walls appears to have stopped and reversed direction. This means that the relatively cold water near the side walls started to convect

upward towards the top layer pushing the relatively warmer water down through the center of the water column. This was reflected in the graph by the sudden cooling of the upper layer while the temperature of the bottom layer remained almost unchanged for a few minutes. Actually, in one of the trials (see Fig 4) the temperature of the bottom layer experienced an increase during this transient period. When the rearrangement of the temperature gradient was completed (colder at the top and warmer at the bottom) the cooling process resumed at a more normal rate. It is interesting to note that the cooling (and also the warming) rate of the water column was not affected by crossing the zero temperature line.

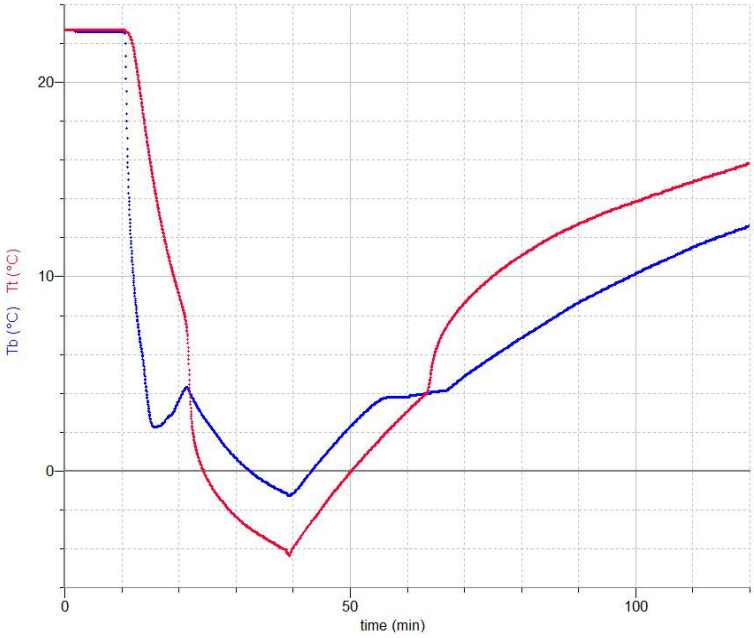


Figure 4: One of the trials of the experiment showing a transitory warming up of the bottom layer near the anomalous temperature during the cooling process of the water column.

When the water column was removed from the cold reservoir and started warming up by the surrounding air at room temperature, the opposite trend in temperature changes was observed. From Fig 2, we see that the temperatures at both ends of the water column initially increased at nearly similar rates. However, when the temperature of the bottom layer approached the anomalous temperature (of 4°C) the warming rate of this layer slowed down very quickly. This was obvious from the corresponding temperature curve which almost reached a plateau that continued for about 13 minutes during which the temperature increased by only a fraction of a degree. This was a

sign that the bottom layer reached maximum density, at this anomalous temperature, and only water at this temperature sank down and remained at the bottom.

The upper layer, on the other hand, continued to warm up at the same rate until it approached the anomalous temperature (see Fig 2). The warming rate, after that, quickly increased crossing the temperature curve of the bottom layer. Apparently, this was triggered by the sinking of the heavy upper layer down the centre of the water column combined with the rise of the relatively less dense warm water near the side walls. When the rearrangement of the density gradient was completed, the water column continued warming up at a normal rate.

From the graphs in Fig 2, we clearly see that the accelerated change in the upper layer's temperature was centered about the 4°C mark. Also, the stabilization of the temperature in the bottom layer is also very close to this temperature within the instrument's precision and accuracy (±0.5°C). To show that such behaviour is a special characteristic of water, a similar procedure can be applied using other liquids (Branca & Soletta, 2005).

6. CONCLUSION

The anomalous behaviour of water near 4°C is a very interesting characteristic of this vital and abundant liquid on our planet. Therefore setting up a lab experiment that clearly demonstrates this behaviour and allows for a quantitative measurement of the anomalous temperature is of great educational value. In this article we presented a simple experimental procedure to achieve this objective, which involve observing the reversal of the temperature gradient in a water column as the temperature changes beyond the anomalous temperature.

The experiment also demonstrated another interesting phenomenon which is the super cooling of water. This demonstration challenges the belief among many students that water must freeze at 0°C.

Finally, we should emphasize that quality physics experiments can be designed at low cost and can be performed safely and independently by the student either as a home lab experiment or as an assignment. We believe that using imagination and dedicated research a wide range of highly quantitative physics lab experiments can be designed that satisfy this purpose. This is especially important in Distance education institutions such as Athabasca University.

REFERENCES

Al-Shamali, F., & Connors, M. (2010). Low Cost Physics Home Lab. In D. Kennepohl & L. Shaw (Eds.), *Teaching Science at a Distance* (pp. 131-146). AU Press.

Branca, M., & Soletta, I. (2005). Thermal Expansion: Using Calculator-Based Laboratory Technology to Observe the Anomalous Behavior of Water. *J. Chem. Educ.* 82, 613-615.

De Paz, M., Pilo, M. & Puppo, G. (1984). Dynamic method to measure the density of water in the vicinity of its maximum. *Am. J. Phys.* 52, 168-171.

Gianino, C. (2007). An easy classroom experiment on the supercooling of water. *Phys. Edu.* 42, 289-292.

Greenslade, T. B., Jr. (1985). The maximum density of water. *Phys. Teach.* 23, 474-477.

Leif, T. R. (2008). Lava Lamp. *Phys. Teach.* 46, 219-221.

Soletta, I., & Branca, M. (2005). The Frozen Lake: A Physical Model Using Calculator-Based Laboratory Technology. *Phys. Teach.* 43, 214-217.

Chapter 3

INTEGRATED MODELING OF THE ATHABASCA RIVER BASIN USING SWAT

Getnet D. Betrie, Baoqing Deng and Junye Wang
Athabasca River Basin Research Institute, Athabasca University

Abstract: The Athabasca River Basin significantly contributes to the provincial economy. There are various sectors such as agriculture, coal mines, oil sands, pulp mills and urbanization. These sectors abstract water and discharge effluent into the basin. The effluents contain excessive nutrients, heavy metals, and organic chemicals that may have adverse environmental and human health risks. In addition, the abstraction of water negatively impacts the environment during the low-flow seasons. An insight into hydrologic processes, contaminant pathways, and mitigation measures through a basin-scale modeling is imperative for sustainable development of the basin. This paper presents database preparation and hydrological modeling of the Athabasca River Basin. The Soil and Water Assessment Tool (SWAT), which is semi-distributed process-based model, was applied to simulate the hydrology process. Digital Elevation Model (DEM), land cover, soil, and climate data were collected and organized into geodatabases using geographical information system (GIS). The geodatabases are extracted with GIS interface and input files to SWAT prepared. Sensitivity parameters were identified and those parameters were calibrated using the observed flow data. The model then validated using separate data to verify its performance. The result shows that the simulated flow matches the observed flow as confirmed by statistical techniques.

Key words: Integrated modelling, Hydrology, Geodatabase, SWAT, Simulations, GIS

1. INTRODUCTION

The Athabasca River Basin, which begins from the Columbia Glacier and drains into Lake Athabasca, significantly contributes to the provincial economy (Athabasca Watershed Council 2011). The main activities that are going in the basin include agriculture, pulp mills, municipal water-treatment plants (WTTP), coal mining, and oil sands mining. Each activity abstracts water and discharges effluent into the basin. The effluents from agriculture, pulp mills, and WTTP contain excessive nutrients such as nitrogen, phosphorus, and organic carbon (Government of Canada, Alberta 2005). The effluents from coal and oil sands mining contain heavy metals and organic chemicals (Government of Canada, Alberta 2005). Excessive amount of nutrients, inorganic, and organic contaminants may have adverse environmental and human health risks. The abstraction of water for purpose of productions also negatively impacts the environment during the low-flow seasons. An insight into hydrologic processes, contaminant pathways, and mitigation measures through a basin-scale modeling is imperative for sustainable development of the basin(Betrie et al. 2011).

Integrated river basin modeling helps decision-makers to understand the basin-scale environmental problems (e.g., point and diffused sources of pollution) by considering the upstream-downstream interdependencies (Betrie et al. 2011). The literature review shows that this approach is widely used to support policy making, particularly in the European Water Framework Directive (WFD) (Grizzetti et al. 2008; Hesse et al. 2008; Volk et al. 2009). Grizzetti et al. (2008) applied the GREEN model to estimate the nitrogen pressures on surface water quality at medium and European scale using readily available data. The results identified the sources and quantity of the diffuse nitrogen emission at the European scale. Hesse et al. (2008) applied SWIM model to identify nutrient pollution, assess the impact of land use and climate changes on water quality and quantity, and evaluate mitigation measures at the Rhine River Basin. The results showed that nitrogen pollution were caused by diffuse sources and could be reduced by application of agricultural measures, whereas phosphorus were caused by point sources and could be reduced by the reduction of point source emissions. Volk et al. (2008) used the SWAT model to evaluate land use and land management scenarios to reduce the total nitrogen concentration in rivers to meet the WFD requirements at the Upper Ems River Basin. The results showed that in order to achieve the requirements of WFD there is a need to reduce arable land, increase pasture and reforestation.

The literature review shows that there is a paucity of research on the impact of the various economic activities have on water quality and quantity of the Athabasca River Basin. Squires et al. (2009) quantified the spatial and

temporal changes of water quantity and quality of the basin between periods 1966-1976 and 1996-2006 using statistical techniques. The results show both the water quality and quantity have changed significantly between the periods. Kelly et al. (2010) conducted field study at the lower part of the basin and reported some heavy metals from oil sands development were exceeded regulatory values near or downstream of development.

The objective of our research team is to develop an integrated modeling framework that can simulate hydrological and biogeochemical processes in the Athabasca River Basin. This framework will enable decision-makers to understand sources and pathways of contaminants and implement best management practices in order to reduce the impact of contaminants in the environment. To this end, this paper presents preliminary results of database development and hydrological modeling of the Athabasca River basin.

2. METHODOLOGY

2.1 Model

Soil and Water Assessment Tool (SWAT) is a semi-distributed model that simulates continuous-time landscape processes at catchment scale (Arnold et al. 1998). The catchment is divided into hydrological response units (HRUs) based on soil type, land use, and slope classes that allows a high level of spatial detail simulation. The major model components include hydrology, weather, soil erosion, nutrients, soil temperature, crop growth, agricultural management, and stream routing.

The model predicts the hydrology at each HRU using the water balance equation, which includes daily precipitation, runoff, evapotranspiration, percolation, and return flow components. The surface runoff is estimated in the model using the Natural Resources Conservation Service Curve Number (CN) method and the Green and Ampt method. The percolation through each soil layer is predicted using storage routing techniques combined with crack-flow model. The evapotranspiration is estimated in SWAT using the Priestley-Taylor, Penman-Monteith and Hargreaves methods.

The SWAT model uses the Modified Universal Soil Loss Equations to compute soil erosion at an HRU level. It uses runoff energy to detach and transport sediment. The sediment routing in the channel consists of channel degradation using stream power and deposition in channel using fall velocity. Channel degradation is adjusted using USLE soil erodibility and channel cover factors.

The amount of nitrogen and phosphorous in water is estimated using load and partitioning concepts (Arnold et al. 1998), respectively. The simulation

of nitrogen and phosphorous cycles are shown in Figure 1. To simulate these nutrients, SWAT takes inputs such as point sources (e.g., industrial effluents) and diffused sources (e.g., fertilizers from agricultural fields). The agriculture management simulates practices like tillage, irrigation, pesticides and fertilizations.

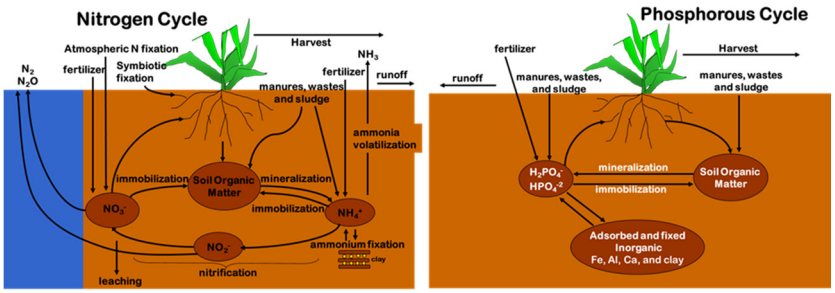


Figure 1: The nitrogen and phosphorous cycles in SWAT (after Arnold et al. 1998)

2.2 Database

The SWAT model requires various inputs that reflect the physiography of a given river basin, which include digital elevation model (DEM), land cover, soil, and weather data. DEM is used to delineate the basin and to extract topographic parameters (e.g., slope and elevation). The delineation process discretizes the basin into hydrologically connected watersheds. The DEM was obtained from the CGIAR-CSI GeoPortal that provides DEM data with 90m spatial resolution and in an ASCII format for the entire world (CGIAR-CSI 2015). A total of six tiles that covers the entire province of Alberta were downloaded. Each DEM was converted into a raster format and combined to form a single map using ArcGIS 10.2 software.

A land cover map is required to identify land covers and estimate their parameters as inputs to the SWAT model. The land cover map was obtained from the GeoGratis portal (GeoGratis 2013). This land cover map has a spatial resolution of 250 m and contains 45 classes that represent vegetative and non-vegetative covers of Canada. A total of 1000 tiles of map were downloaded in a vector format. These maps were combined, converted into raster format, and projected into NAD 183Transverse Mercator using Arc GIS 10.2 software. Then the land cover map of Alberta was extracted from the combined map. A database that stores parameters (e.g., temperature responses, leaf area development, residue decomposition, and others) of each land cover was built based on the available SWAT crop database.

A soil map is required to identify various soil types and parametrize their physical and chemical characteristics as inputs to the model. The soil data was obtained from Agriculture and Agri-Food Canada (Agriculture and Agri-Food Canada 2013). The data consists of polygon attribute table (PAT), soil name table (SNT), and soil layer table (SLT). The PAT contains attributes such as polygon area, perimeter and their polygon id. The SNT describes the physical and chemical parameters of soils that are stored in in the PAT. It also describes the soil order and groups. The SLT contains information which varies in a vertical direction for each soil stored in the PAT. The collected data in a shapefile format are imported to ArcGIS 10.2, their attribute tables joined, and converted into raster format, and its spatial reference corrected to Transverse Mercator. The final soil map covers the entire province of Alberta, has 2153 soil types, and a spatial resolution of 1 Km. A database that stores all soil types and their associated physical and chemical characteristics was built based on the available SWAT database. The physical properties (e.g., hydrologic group, texture, available water holding capacity, soil erodability constant, bulk density and other) of the soil govern the movement of water and air the profile. On the other hand, the chemical properties (e.g., anion exchange capacity, organic carbon content, pH, EC and other) are used to setup initial amount various chemical in the soil.

Climate data such as precipitation, maximum/minimum air temperature, solar radiation, wind speed and relative humidity are required in order to simulate the hydrological processes. Daily climate data were collected from 126 stations that are located in the river basin. The data was obtained from Environment Canada portal (Environment Canada 2015). These stations do have solar radiation data, relative humidity and have many missing values. Precipitation and temperature inputs for SWAT were prepared for 50 climate stations out of 126 stations. The missing values in precipitation and temperature data were filled using a spatial interpolation technique. The other climate variables were estimated from global weather database that was obtained from NCEP climate forecast system reanalysis (Saha et al. 2010).

2.3 Model Setup

The ArcGIS interface (Winchell et al. 2013) was used to delineate the basin and prepare input files for the SWAT model. The basin was divided into 9 watersheds by defining a threshold area of 600,000 ha, which is the minimum upstream drainage area required to define the beginning of a stream. Figure 2 shows the Athabasca river basin, its watersheds, and the river and its tributaries. In this model, the nine watersheds are shown as

polygons. For each watershed, its slope, elevation, area and geographical location are estimated and stored in topographic database. The blue points in the map are the outlets of each watershed and the basin.

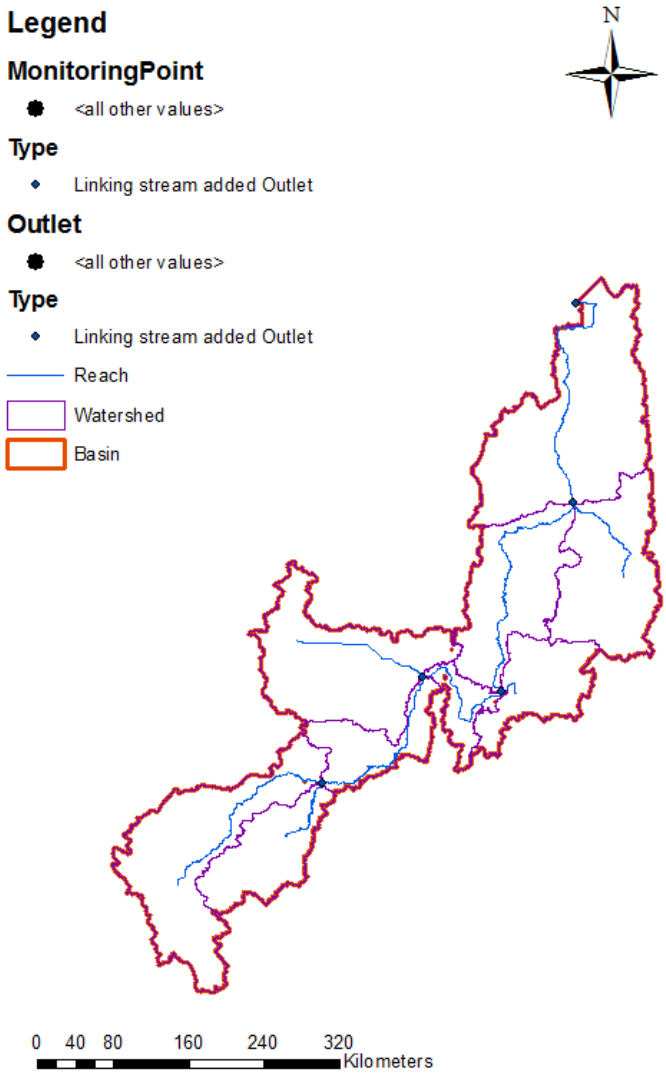


Figure 2: The Athabasca River Basin and its watersheds

Figure 3 depicts the land cover and soil classes for the Athabasca River Basin. The model estimated 11 cover classes from the provided land cover map. These cover classes are water, barren land, shrub land, forested wetlands, non-forested wetland, herbaceous land, range land, agriculture land, pasture land, evergreen forest, and mixed forest. In addition, the model estimated 119 different soil types from the provided soil map. The dominant

soil types at the upper, central, and lower parts of the basin are Hubalta, Kinosis, and Fire Bag, respectively.

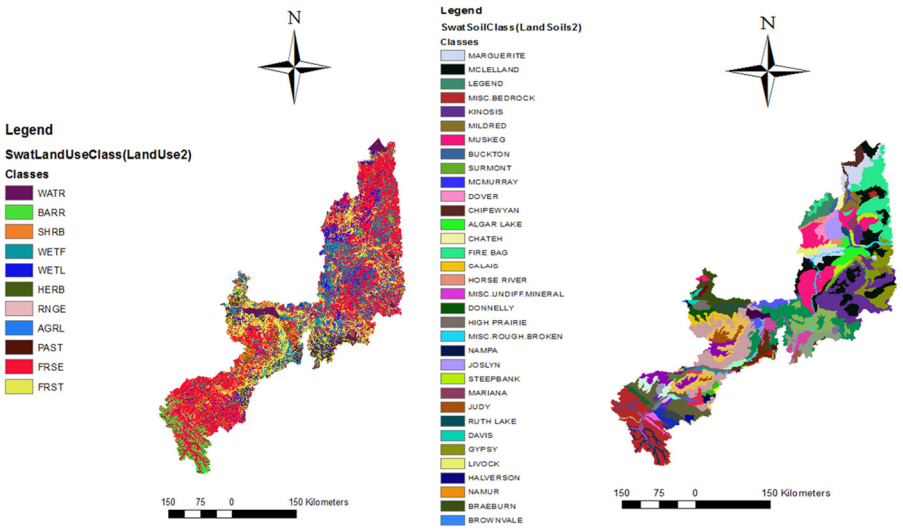


Figure 3: Land cover (left) and soil (right) classes of the Athabasca River Basin

Figure 4 shows the slope classes of the Athabasca River Basin. The slopes are computed from the provided DEM. In this model, five slopes classes are computed, which are 0-5%, 5-15%, 15-30%, 30-50%, and greater than 50%. The upper part of the basin is characterized by greater than 30% slope. The central part of the basin is characterized by slope between 5-15%. The lower part is characterized with slope less than 5%.

The land cover, soil, and slope maps were overlaid to derive 1362 dominant HRUs. HRU is a unique combination of soil, land cover, and slope class. These HRUs are used by the model to simulate hydrological and water quality processes in the basin. The model was run for 15 years from 1990 to 2004. The first three years were used to warmup the model, while the periods 1993 to 1998 and 1999 to 2004 were used to calibrate and validate the model, respectively. SWAT has over 300 parameters, but a few parameters must be selected for simulation to reduce the model complexity. Thus, sensitivity analysis was conducted to identify most sensitive parameters for model calibration using the One-factor-At-a-Time algorithm (van Griensven et al. 2006). The sensitive parameters were calibrated using Sequential Uncertainty Fitting algorithm (Abbaspour et al. 2007). The performance of the SWAT simulation was measured using statistical techniques, which are R2 and Nash-Sutcliffe efficiency (NSE).

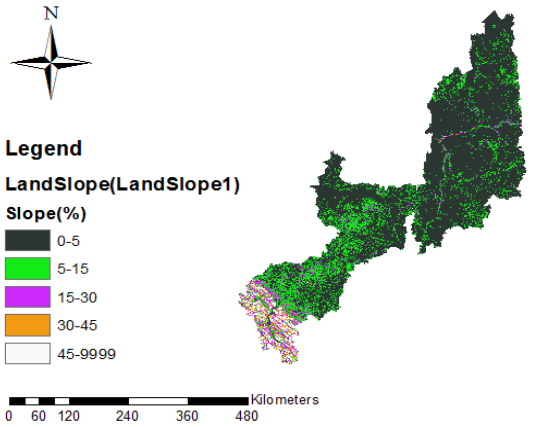


Figure 4: Slope class of the Athabasca River Basin

3. RESULTS

The results of sensitivity analysis are shown in Table 1. The most sensitive parameters are snow melt base temperature (SMTMP), snowfall temperature (SFTMP), maximum melt factor for snow (SMFMX), snowmelt temperature lag factor (TIMP), curve number (CN2), surface runoff lag time (SURLAG), base flow alpha factor (ALPHA_BF), groundwater delay time, (GW_DELAY), recharge to deep aquifer (RCHRG_DP), soil evaporation compensation factor (ESCO), plant uptake compensation factor (EPCO), and minimum melt factor for snow (SMFMN). These parameters with their fitted values are used to simulate the hydrology of the Athabasca River Basin.

Table 1. The sensitive parameters, their range and fitted values

Parameter	Description	Minimum value	Maximum value	Fitted value
SMTMP.bsn	Snow melt base temperature	0	5	0.5
SFTMP.bsn	Snowfall temperature	-10	5	3.5
TIMP.bsn	Snowmelt temperature lag factor	0	1	0.7
SMFMX.bsn	Maximum melt factor for snow	0	5	1.5
CN2.mgt	Curve number	0	1	0.3
SURLAG.bsn	Surface runoff lag time	0.05	2	0.245
ALPHA_BF.gw	Base flow alpha factor	0	1	0.9
GW_DELAY.gw	Groundwater delay time	0	1	0.3
RCHRG_DP.gw	Recharge to deep aquifer	0	1	0.9
ESCO.hru	Soil evaporation compensation	1	1	1

	factor			
EPCO.hru	Plant uptake compensation factor	0	3	2.7
SMFMN.bsn	Minimum melt factor for snow	0	3	0.3

The flow simulation of SWAT was calibrated from 1993 to 1998 and the result is shown in Figure 4. The simulated flow matched the observed flow by R2 is equal to 0.61 and NS is equal to 0.32. In both R2 and NSE measurements, the model is acceptable but not that much good. The simulated flow overestimates the observed flow during the peak flow. The model well simulated the low flow except for the years 1997 and 1998. The simulation of the model well captured the rising and falling limbs of the observed hydrograph.

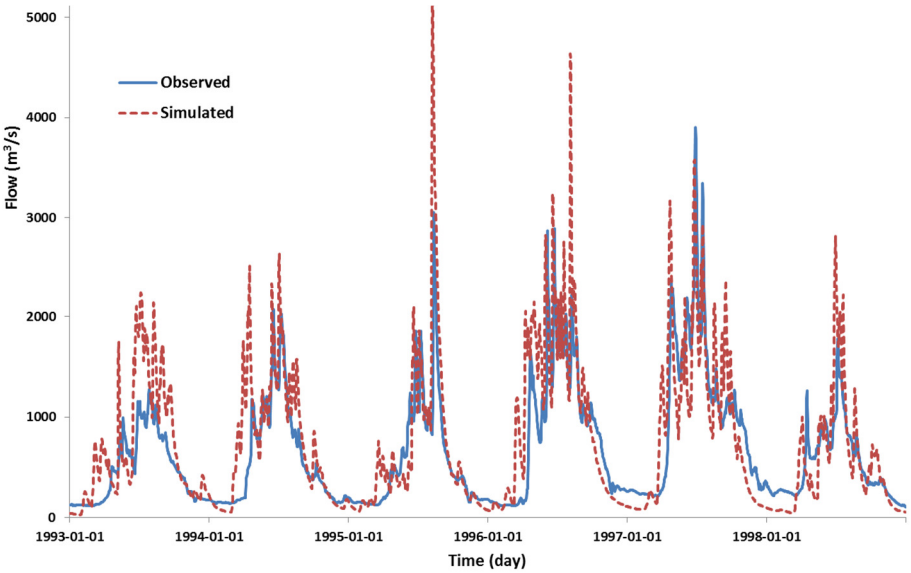


Figure 5: Simulated and observed flows for the calibration period

The flow simulation of SWAT was validated from 1999 to 2004 with the observed dataset that was not used during the calibration period. The result of the flow validation is presented in Figure 6. The simulated flow matched the observed flow by R2 and NSE values are equal to 0.65 and 0.5. This indicates that the model is quite acceptable in both R2 and NSE measurements. SWAT well simulated the observed flow during the peak flow except the summer of 2002. However, the model underestimated the observed flow during the low flow that occurs during winter. In most of the weather stations that are located in the basin, there are no measurements of

climatic data. This could be the main reason for the discrepancy between the simulated and observed flow during winter.

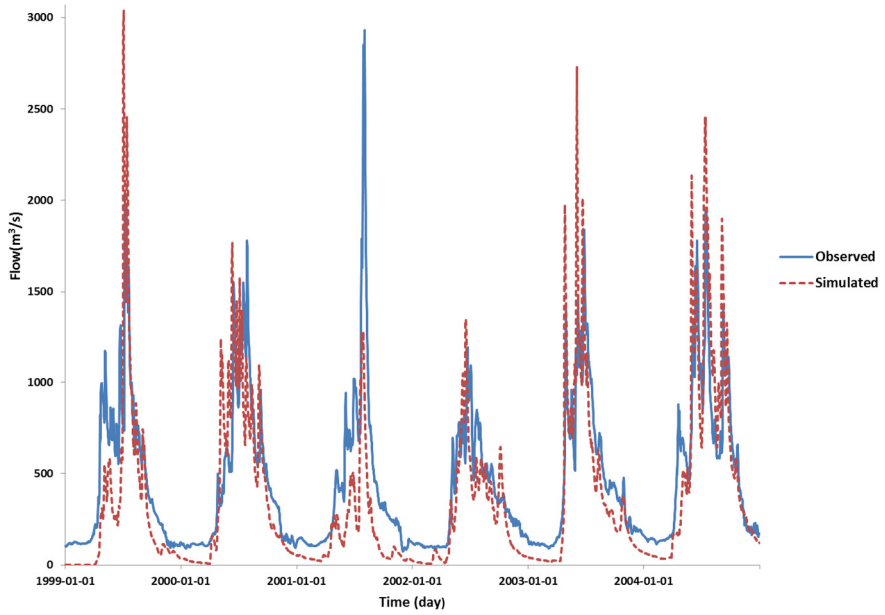


Figure 6: Simulated and observed flows for the validation period

4. CONCLUSION

This paper presents database preparation and hydrological modeling of the Athabasca River Basin. Data such as DEM, land cover, soil and climate data were collected and organized into geodatabases using geographical information system. The prepared databases are used as input to the SWAT model to simulate the hydrology of the basin. Sensitive parameters were identified and calibrated using automatic sensitivity and calibration tools. The performance of the model was evaluated using statistical techniques.

The result of the model validation shows that the simulated flow is comparable to the observed flow. However, the model systematically underestimates and overestimates the low and peak flows, respectively. This is most likely attributed to poor quality of climatic data used as an input to the model, which have huge missing values. Further work is being done to improve the limitation of this model. In conclusion, the preliminary result indicates that SWAT can be applied to understand the hydrological processes of the Athabasca River Basin.

REFERENCE

Abbaspour, K.C. et al., 2007. Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. Journal of Hydrology, 333, pp.413–430.

Agriculture and Agri-Food Canada, 2013. The national soil database (NSDB). Available at: http://sis.agr.gc.ca/cansis/nsdb/slc/v3.2/index.html [Accessed March 21, 2015].

Arnold, J.G. et al., 1998. LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART I : MODEL DEVELOPMENT. , 34(1), pp.73–89.

Athabasca Watershed Council, 2011. Athabasca Watershed Council state of the watershed report: Phase I,

Betrie, G.D. et al., 2011. Linking SWAT and SOBEK Using Open Modeling Interface (OPENMI) for Sediment Transport Simulation in the Blue Nile River Basin. American Society of Agricultural and Biological Engineers, 54(5), pp.1749–1757.

Betrie, G.D. et al., 2011. Sediment management modelling in the Blue Nile Basin using SWAT model. Hydrology and Earth System Sciences, 15(3), pp.807–818. Available at: http://www.hydrol-earth-syst-sci.net/15/807/2011/ [Accessed May 23, 2013].

CGIAR-CSI, 2015. CGIAR-CSI SRTM 90m DEM Digital Elevation Database. Available at: http://srtm.csi.cgiar.org/ [Accessed January 5, 2015].

Environment Canada, 2015. Historical climate data. Available at: http://climate.weather.gc.ca/ [Accessed January 21, 2015].

GeoGratis, 2013. GeoBase land cover map. Available at: http://geogratis.gc.ca/api/en/nrcan-rncan/ess-sst/$categories?scheme=urn%3Aiso%3Aseries&q=GeoBase [Accessed March 21, 2015].

Government of Canada, Alberta, N.T., 2005. Northern Rivers Ecosystem Initiative: Synthesis Report, Available at: http://www.mb.ec.gc.ca/nature/ecosystems/nrei-iern/index.en.html.

Van Griensven, a. et al., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. Journal of Hydrology, 324, pp.10–23.

Grizzetti, B., Bouraoui, F. & De Marsily, G., 2008. Assessing nitrogen pressures on European surface water. Global Biogeochemical Cycles, 22(August 2007).

Hesse, C. et al., 2008. Eco-hydrological modelling in a highly regulated lowland catchment to find measures for improving water quality. Ecological Modelling, 218, pp.135–148.

Kelly, E.N. et al., 2010. Oil sands development contributes elements toxic at low concentrations to the Athabasca River and its tributaries. Proceedings of the National Academy of Sciences of the United States of America, 107, pp.16178–16183.

Saha, S. et al., 2010. The NCEP climate forecast system reanalysis. Bulletin of the American Meteorological Society, 91(August), pp.1015–1057.

Squires, A.J., Westbrook, C.J. & Dube, M.G., 2009. An approach for assessing cumulative effects in a model river, the Athabasca River Basin. Integrated environmental assessment and management, 6(1), pp.119–134.

Volk, M., Liersch, S. & Schmidt, G., 2009. Towards the implementation of the European Water Framework Directive?. Lessons learned from water quality simulations in an agricultural watershed. Land Use Policy, 26, pp.580–588.

Winchell, M. et al., 2013. ArcSWAT interface for SWAT2012: User's guide. , p.464.

ACKNOWLEDGEMENTS

The authors gratefully acknowlede the support provided for this research by the Campus of Alberta Innovates Research Chair RCP-12-001-BCAIP.

Chapter 4

IMPROVING LEARNING BASED ON THE IDENTIFICATION OF WORKING MEMORY CAPACITY, ADAPTIVE CONTEXT SYSTEMS, COLLABORATIVE LEARNING AND LEARNING ANALYTICS

Richard A.W. Tortorella[1], Darin Hobbs[2], Jeff Kurcz[2], Jason Bernard[2], Silvia Baldiris[2,3], Ting-Wen Chang[4], Sabine Graf[2]
[1] *University of Eastern Finland, Finland*
[2] *Athabasca University, Canada*
[3] *Fundación Universitaria Tecnológico Comfenalco, Colombia*
[4] *Beijing Normal University, China*

Abstract: Working memory capacity and learning styles play key roles within adaptive learning environments. In addition, the concepts of collaborative efforts, context awareness, ensuring student engagement and the identification of students at risk of dropping out, play vital roles and are key to any successful learning environment. In this chapter, key concepts and mechanisms for each of them are discussed along with various approaches and frameworks. A means of utilizing artificial intelligence to improve working memory capacity identification and learning styles identification is discussed in the second section. Adaptation is discussed in both the third and fourth section, as it pertains to collaborative learning environments and adaptive context-aware expert systems. The final two sections address the problem of student drop-out rates as it pertains to improving the promotion of scientific competencies and the identification of students at risk of dropping out. All these concepts assist in providing learners with adaptive and improved learning environments that aid in supporting learners in the learning process.

Key words: learning style, working memory capacity, collaborative learning, context aware, scientific competences, learning analytics

1. INTRODUCTION

There is no single definition or template of a learner. Learners come from different backgrounds and have different skillsets, motivations environments and problems. The aim of our research is to facilitate the creation of more effective, adaptive and tailored learning systems that are conducive to learning at every stage of the learning process. In this chapter, we focus on five key areas and present research conducted in these areas: (1) the identification of working memory capacity and learning styles, (2) adaptive recommendations for collaborative learning, (3) adaptation based on context information, (4) adaptive recommendations for learning scientific competencies and (5) identification of students at risk of dropping out or failing.

Working memory capacity (WMC) play a key role is affecting a student's behaviour on how they perform in reading comprehension, decision making and problem solving (Ford & Chen, 2001; Broadway & Engle, 2011). The careful consideration of a student's WMC assists in the prevention of cognitive overload and thus able to affect student's learning in a positive manner (Gathercole & Alloway, 2008). The approach described in Section 2 of this chapter suggests a means of improving precision of learning style and WMC identification. This would permit students to benefit through more appropriate content matching or better advice from their teachers.

Collaborative learning is beneficial and crucial to the learning process (Koh, Barbour & Hill, 2010). In Section 3 we address the concept that collaborative learning can be enhanced by utilizing computer support. The research discussed in this section aims to provide adaptive recommendations to collaborative groups while they are working on a group project, focusing on support for project management and communication aspects.

Section 4 addresses the concepts of adaptive context-aware expert systems. Environmental and spatial conditions can have a significant impact on the way we learn. Although learning experiences, thanks to mobile technology, are able to take place anytime and anywhere (Shih et al., 2010), the increasing variety of locations and conditions where learning can occur has led to serious technical and contextual challenges. The research in this section describes the development of a framework that would facilitate the creation of adaptive context-aware systems that integrates with an adaptive engine.

The acquisition of scientific competences is a key issue in postgraduate programs. Section 5 suggests a solution to improve the experience of beginning researchers when they do research through the generation of recommendation in each step of the research process by the use of an ontology that represents practical and conceptual knowledge about research

methods. The generated recommendations facilitate the decision-making process in the research process.

According to McNutt and Brennan (McNutt & Brennan, 2005) reports published in 2005, in the Chronicle of Higher Education (US) have found that post-secondary institutions are seeing dropout rates ranging anywhere from 20% to 50% for distance learners. Section 6 demonstrates means of determining variables that are the most relevant in the successful identification of students at risk.

2. IMPROVING PRECISION OF LEARNING STYLE AND WORKING MEMORY CAPACITY IDENTIFICATION WITH ARTIFICIAL INTELLIGENCE

The identification of students' learning styles, their preferences towards the learning task, and working memory capacity (WMC), the number of items they can store in short term memory, allows personalized content to be matched to the student. The student benefits from learning style identification with improved learning outcomes (e.g., Ford & Chen, 2001), satisfaction (e.g., Popescu, 2010), and a reduction in learning time (e.g., Graf, Chung, Liu & Kinshuk, 2009). Although there are many models for learning styles, this research used the Felder-Silverman learning style model (Felder & Silverman, 1988) which consists of four dimensions: active / reflective (A/R), sensing / intuitive (S/I), visual / verbal (V/V) and sequential / global (S/G).

Questionnaires exist which can identify students' learning styles and WMC; however, these have two notable drawbacks. Questionnaires are intrusive to the learning process. Also, questionnaires may be influenced by other factors such as a student's mood, so some students' characteristics will not be accurately identified. Automated approaches overcome intrusiveness by working in the background and by using students' behaviors they are less subject to other factors. The drawback to automated approaches is that they peak at about 80% precision leaving some room for improvement. This research aims to answer how artificial intelligence can be used to improve precision of automated approaches while being general to any learning management system.

One approach, DeLeS (Graf, Kinshuk & Liu, 2009; Chang, El-Bishouty Graf & Kinshuk, 2013), uses behavior patterns which are general to any learning management system to identify learning style and WMC and has a

leading degree of precision (~80%). One issue with DeLeS is that it assumes that all behavior patterns are equally important by assigning each behavior pattern a weight of 1. If an optimal set of weights could be found then precision should be increased; however, the set of all weight combinations is a very large space (minimum 10^{12} combinations) and so three optimization algorithms are proposed to search more efficiently. Alternatively, the behavior patterns may serve as good inputs directly into a classification algorithm.

The three optimization algorithms selected were ant colony system (ACS), genetic algorithm (GA) and particle swarm optimization as each explores the solution space in a different manner and thus may give different results. The classification algorithm selected was the artificial neural network (ANN). Each of these approaches was named LSID (Learning Style Identifier) with the corresponding algorithm, for example LSID-GA for genetic algorithm. To assess all four approaches, 75 students' behaviour and learning style data and 63 students' behavior and WMC data was used. For each approach, the remainder of this process was repeated for each learning style dimension and WMC. Each algorithm has several parameters which greatly influence the ability of the algorithm to be trained properly. To optimize the parameters, each parameter was systematically altered one a time within ranges suggested from literature. Overfitting is a common problem with artificial intelligence algorithms where solutions are fit to the data's noise, so next overfitting reduction techniques were assessed. For all four algorithms stratification (Kohavi, 1995) was assessed and for the ANN future error prediction (Mitchell, 1997) and weight decay (Krogh & Hertz, 1992) were assessed additionally. To further promote a general nature to the algorithms, a 10 fold cross validation technique was used for every execution thus ensuring that the algorithms are able to work under a variety of data sets. For this reason, all the results are averages over the 10 folds.

With the optimal parameter settings and optimal overfitting reduction techniques, a final result was obtained shown in Table 1. These results show that, with two exceptions, all of the LSID approaches improve precision over DeLeS. The exceptions are LSID-GA in the A/R dimension which is worse than DeLeS and LSID-ANN in the S/I dimension which is equal to DeLeS. The best results were obtained by the ACS in the A/R and S/I dimensions while the ANN obtained the best results for V/V, S/G dimensions and for WMC. These results show that precision is improved by finding an optimal set of weights for the behavior patterns and that these behaviour patterns are successful as direct inputs into a classification algorithm.

Table 1. Comparison of precision results for LSID and DeLeS (ranking in parenthesis, and top result bolded

Approach	**A/R**	**S/I**	**V/V**	**S/G**	**WMC**

LSID-ACS	**0.819 (1)**	**0.797 (1)**	0.799 (2)	0.737 (4)	0.855 (2)
LSID-GA	0.795 (5)	0.796 (2)	0.794 (4)	0.774 (2)	0.836 (3)
LSID-PSO	0.805 (2)	0.794 (3)	0.796 (3)	0.768 (3)	0.835 (4)
LSID-ANN	0.802 (3)	0.790 (4)	**0.840 (1)**	**0.797 (1)**	**0.862 (1)**
DeLeS	0.799 (4)	0.790 (4)	0.788 (5)	0.702 (5)	0.809 (5)

By improving precision of learning style and WMC identification students would be expected to benefit through more appropriate content matching or better advice from their teachers. This would then lead to improved learning outcomes and satisfaction, and a reduction in the time needed to learn.

3. PROVIDING SUPPORT FOR STUDENTS THROUGH ADAPTIVE COLLABORATIVE LEARNING ENVIRONMENTS

Collaborative learning is an important aspect of the learning process. Learning goes beyond the learning material to teach students and builds upon other important indirect skills such as communication and interpersonal skills (Williams & Roberts, 2002). Collaborative learning can be greatly enhanced by the use of computer supported collaborative learning (CSCL) if it is properly implemented, especially when an adaptive learning system assists students along the way. Implementing a collaborative learning environment is both beneficial and crucial to the learning process if used correctly (Koh, Barbour & Hill, 2010). Learning management systems (LMS) are becoming widely popular with schools and education systems for both online and blended learning environments and while these systems are great for presenting information to students, they do not necessarily support collaborative learning or provide intelligent features to facilitate the collaborative learning process.

Our research aims to deliver adaptive recommendations to students working in groups in order to benefit and assist their progress. In this section, we introduce the Adaptive Collaborative Systems (ACS) which supports collaborative learning in learning management systems, focusing on two areas: communication and project management. ACS is different in terms of how other system have been designed and implemented as it is presenting adaptive recommendations to students and/or groups as they work on team projects. Other differences include that ACS is domain independent and the design can be integrated into any LMS.

3.1 Communication features in ACS

ACS monitors the communication between students to determine participation in a way that can ensure students are actively attending meetings as well as contributing to dialogues fairly. ACS looks at both ends of the spectrum of communication, high and low, in a variety of communication channels, including forums, chats and imported messages (e.g., from Skype). Brindley, Walti, & Blaschke (2009) discussed that not all communication will take place on the LMS and that other third party applications may be used. Therefore our system incorporates a utility that allows students to import third party chat logs for both participation analysis and centralized log to reference at later if needed.

If ACS determines that a student is not attending meetings frequently, it provides this student with an alert. Furthermore, if ACS determines that a student contributes significantly less than others to a dialog, ACS alerts the student privately to participate and contribute more to the dialog. ACS also uses the same monitoring techniques to determine if students are over participating and recommends those active users to improve their leadership skills by include other quieter members into the conversation, for example, by asking them direct questions about the content so they can elaborate further.

3.2 Project management features in ACS

ACS takes advantage of the information provided by the students on their progress for each task. This allows the system to help keep students on task in a timely manner by monitoring the amount of work done for each task and comparing it to the time that has passed. This information is presented to

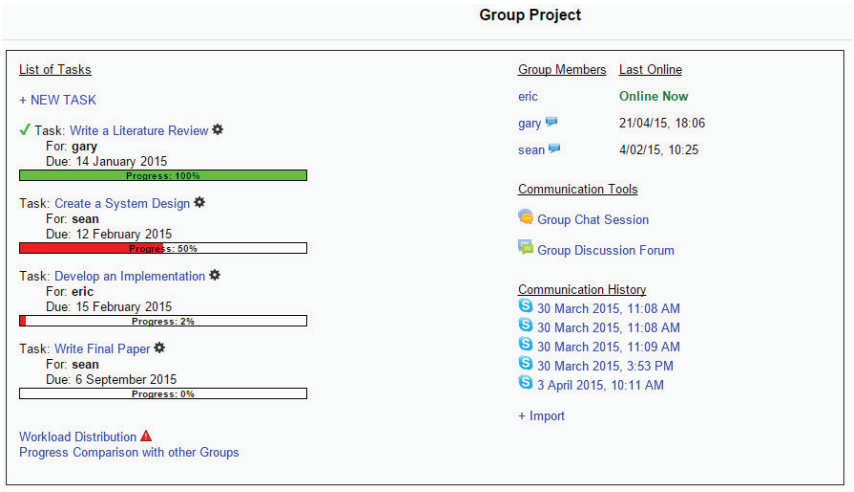


Figure 1. Main interface of ACS

students using an at-a-glance interface (see Fig. 1) so students can quickly and easily tell if they are on track on a specific task or not by displaying a progress bar in green or red, respectively. To motivate students, ACS also displays a group's current progress in comparison to other groups' progress in the current course. Furthermore, it monitors a group's risk of failing by comparing their progress to previous cohorts' progress at the same time and alerts a group if they become at risk of failing. ACS also ensures that an equal distribution of work has been assigned to all students by comparing the allotted amount of work between all group members.

Especially in online learning settings, group work is used more and more due to its many advantages and ease of use. ACS aims at ensuring that the potential of collaborative learning can be fully unlocked and all students are supported to participate and benefit equally from collaborative learning through personalized recommendations and information on how to effectively learn in collaborative settings.

4. A GENERIC PLATFORM FOR ADAPTIVE CONTEXT-AWARE EXPERT SYSTEMS

Collaborative learning has been seen as an important part of the learning process. It can take place over short distances in close proximity or with modern technology, between two places anywhere on the globe. Yet, our surroundings affect almost every aspect of our daily lives: from the mundane to the most elaborate task, where we are and what our proximate environment is like, affects us in many varied ways. This is also true for our ability to learn, specifically how and what information is presented to us. With the ubiquitous nature of smartphones worldwide, researchers are being provided enormous processing power in the hands of learners.

Although it has a great effect on us, people may not be always cognizant of their surroundings or its effect on their daily lives. For example, Schilit, Adams, and Want (1994) described a system that reacted to an individual's changing context. They stressed the importance of the limited information within a person's proximate environment.

Although proposed over 20 years ago, it is only recently thanks to technological advances that we are able to adequately investigate and apply this type of system as proposed by Schilit, Adams, and Want (1994) to the general public, specifically to aid in learning. Much work has been done since the 1990s to further the field, however there has not been much research done towards the creation of a framework that would incorporate

context aware adaptive learning systems. For example, Anagnostopoulos and Hadjiefthymiades (2009) described an extension of context presentation that would help in the representation, classification and inference of sensor data obtained from a device. In the educational domain, Liu and Hwang (2010) described the paradigm shift between conventional e-learning to m-learning to context-aware ubiquitous learning.

This has extensive ramifications into learning in general, as environmental conditions can affect the way we learn, and the type of information we require changes with our environment. This raises the concern of how to provide advanced adaptive learning based on environmental and contextual information. In order to help find a solution for this issue, our research involves the integration of an adaptive context-aware learning system and an expert system.

As many current context-aware systems are designed to work with specific scenarios, when a different scenario is needed, the system typically needs to be re-built from scratch. Our research proposes a generic framework that integrates the inference engine of an expert system with a context-aware, mobile adaptive engine. This research aims to answer the following questions: How to automatically detect context information and create a generic rich context model for adapting to context and environmental factors? Furthermore, how does an adaptive context-aware system integrate with the inference engine of an expert system in order to allow for a generic platform between the two systems?

Figure 2 shows a brief overview of the main components of the framework and their basic functionality.

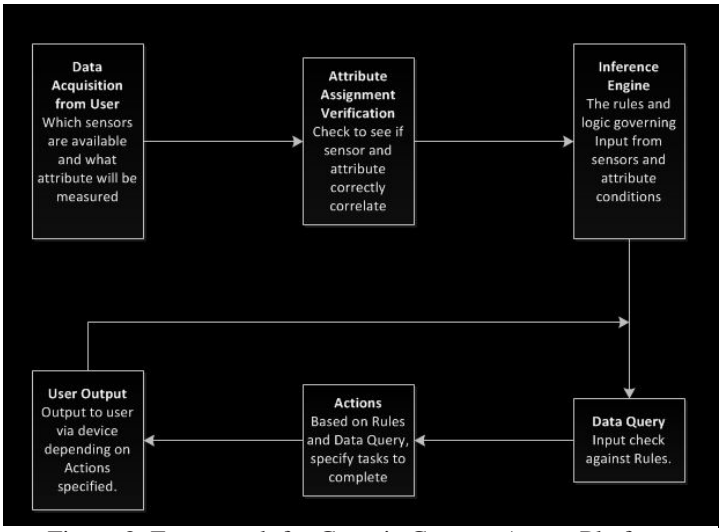


Figure 2. Framework for Generic Context-Aware Platform

The outcome of this research will be a generic platform that after minor configurations can be adapted to many different types of knowledge bases and inference rules, and therefore would be applicable in different scenarios.

5. IMPROVING THE EXPERIENCE OF BEGINNING RESEARCHERS WHEN THEY DO RESEARCH

How to promote the acquisition of scientific competences (NPA Core Competencies Committee, 2009) is one of the most important issues and permanent and persistent problems through the years in the context of postgraduate programmes. It is a relevant problem for postgraduate programmes because it causes reduction in the quality of postgraduate research as well as high drop-out and late submission rates. Important studies have been conducted with the purpose to identify causes of a very common problem that face beginning researchers as well as the attitude of postgraduate students toward research (Graves, 1976; Reeves, 2000; Shaukat, Siddiquah, & Abiodullah, 2014; Zuber-Skerritt, 1987). In general detected problems could be summarized in the following categories:

- Inadequate supervision
- Emotional and psychological problems
- Lack of understanding and communication between supervisor and student
- Student's lack of the fundamentals of scholarship due to a lack of background knowledge, training or experience in research methods
- Late completion and high drop-out rates

The origin of these problems arise in that it is "often assumed in postgraduate education that candidates have developed basic research and writing skills at undergraduate level (reading, note-taking, essay writing, problem solving, information and retrieval skills, etc.) and they are able to translate and apply these skills to their thesis research and writing…" (Zuber-Skerritt, 1987).

On the other hand, the generalized adoption of the single-supervisor model of postgraduate teaching that indicate "…whether and how well a student is guided in the research process and helped in developing skills in thesis writing, depends solely on the individual supervisor's available time, attitude and ability to teach these skills" (Zuber-Skerritt, 1987). In this way, if the supervisor does not have time or the necessary knowledge to support

the student then, the research process could be unsuccessful, which frequently happens.

Several solutions that come from educational perspective have been proposed and validated to alleviate the described problem:

- The reviews of postgraduate programs to include educational strategies centered in the students' needs and preferences.
- The introduction of a workshop model for developing skills in dissertation research and writing.
- Many courses about challenges and methods in research have been created

Our hypothesis is that it is possible to conduce beginning researchers to a successful research if they receive appropriate recommendations, including conceptual and practical ones, based on the practical and conceptual knowledge about research methods represented in an ontology, that helps then to take high quality decision in each step of the research process.

Our solution includes:

- The generation of an ontology that represents practical and conceptual knowledge about research methods;
- The design of a recommender system that uses the generated ontology to give answers to typical problems or questions faced by beginning researchers.
- The evaluation of the system through an experiment with real students.

As recommended on by Shaukat et al. (2014) it is important nowadays to develop positive attitudes in the students toward research. Our research aims at contributing towards this goal by reducing the uncertainty that beginning researchers face when they do research.

6. RELEVANT VARIABLES FOR IDENTIFYING STUDENTS AT RISK

Several studies have shown that online distance students often have higher dropout and failure rates than students who attend classes in a physical environment (Lokken & Mullins, 2014; Schaeffer & Konetes, 2010). There are a number of aspects as to why online students do not succeed as frequently as their offline counterparts. Factors can include feelings of isolation, dissociation with the learning environment, differing learning styles, and technical difficulties (Schaeffer & Konetes, 2010). One of the major challenges of ensuring student success in an online environment is the lack of direct, face-to-face contact with instructors and other students.

Consequently, instructors are unable to observe when a student becomes disengaged or distracted from the learning process.

Fortunately, in an online environment the primary method of course delivery is through a learning management system (LMS), where activities from students, teachers, and course administrators, are automatically captured in database tables or log files (Mazza & Dimitrova, 2004). However, the online activity reporting functionality from an LMS is limited, primarily providing simple reports, including, for example, the last logins of students or the total number of logins for a specified date range. As with other organizations across industries, educational institutions have found themselves in the 'data rich, information poor' paradox.

Institutions have been employing learning analytics, which is the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Siemens, 2010), to develop prediction, risk identification, and intervention systems to increase student success (Chatti et. al, 2012). However, predicting student success at one institution is not guaranteed for predicting the success of students at other institutions. As more educational institutions and LMS vendors develop more learning analytic-style analysis and reporting tools, the field of learning analytics can benefit from research in determining variables that are the most relevant in the successful identification of students at risk.

6.1 Determining Relevant Variables

The intent of our study is to assist in furthering the development of the learning analytics research field by determining the relevancy of variables in the identification of students at risk. The three research questions that guided the direction of this study are:

1. What variables have been relevant in past studies?
2. Based on real student data provided for this study, which variables are accurate at identifying students at risk?
3. Based on real student data provided for this study, which variables are more relevant than others?

In addition to an online search of academic and scientific publications, this study included the analysis of student background and behavioral data over a period of five years. Past research has applied differing data mining or statistical analysis techniques on student behavioral data from a single institution, on a small number of courses, for a short period of time, typically one or two semesters, or for a specific student population, such as first year

students or some combination of this sample. The data used in our study spans over sixteen semesters, nine courses, 94 classes, and 320 students at various points in their academic studies. The courses were fully dependent on internet and communication technologies for course delivery.

6.2 Methodology

To answer Question #1, what variables have been relevant in past studies, an online search and literary review of a number of academic and scientific publications was conducted. The online search consisted of the following key words and phrases: *learning analytics*, *educational data mining*, *academic* and *action analytics*, *student success* and *retention*, *predictive modeling software*, *tools*, and *online* and *distance education*.

To answer Questions #2, based on the student data provided for this study, which variables are accurate at identifying students at risk, and #3, based on the student data provided for this study, which variables are more relevant than others, student background and behavioral data from nine courses in a computer science graduate program, from the Fall 2007 semester to the Spring 2012 semester, was obtained. In total, there were 320 students comprising 1300 records, including their final course grades.

Two methods, Spearman's correlation coefficient for independent to dependent variable analysis and Pearson's correlation coefficient for independent to independent variable analysis was applied to partially answer Questions #2 and #3. Additional analysis was conducted by creating a Bayesian network for each course (i.e., each course revision). Descriptive statistics were used to determine the frequency, significance and direct relationships between all variables, and ultimately identify the variables that were the most relevant in terms of significance, strength and frequency of relationships.

6.3 Results

To answer Question #1, what variables have been relevant in past studies, over 200 variables from 22 studies were used in ensemble models to identify students being at risk. The number of discussion postings created was the prediction variable that was reported to be significantly relevant the most (61.5%) frequent in empirical research. A listing of prediction variables used in more than ten studies and the percentage of studies in which the variable was found to be relevant is displayed in Table 2.

Table 2 – Prediction variables appearing in ten or more studies and their reported relevancy percentage

Prediction Variables	Number of Studies Variable Appeared	Percentage of Studies Variable Reported as Relevant
Number of Discussion Postings Created	13	61.5%
Assessment Tests	13	53.9%
Number of Discussion Postings Viewed	11	45.4%
Number of Resources Viewed	22	40.9%
Grade Point Average	11	27.3%
Number of Mail Messages Created	11	27.3%

To answer Question #2, based on the student data provided for this study, which variables are accurate at identifying students at risk, the results of the correlation analysis and the number of direct relationships with independent variables to the dependent variable were combined. The variables for the total number of files uploaded to the LMS (average p value = .004, average r_s = .409) and total number of discussions viewed (average p value = .007, average r_s = .318) are identified as being significantly related to the final grade the most, appearing in 75% of the data sets, as displayed in **Error! Reference source not found.**3.

Table 3 - Correlation analysis results of relevant prediction variables with student final grade, including the percentage that direct relationships existed in the course data sets

Prediction Variables	p	r_s	Direct Relationships in Course Data Sets	Course Data Sets Appearance
Total Files Uploaded	0.004	0.409	56.3%	75.0%
Total Discussion Posts Viewed	0.007	0.318	68.8%	75.0%
Total Number of Log Activity Records	0.007	0.400	75.0%	68.8%
Percent of All Post Types	0.003	0.391	37.5%	68.8%
Average Number of Discussion Posts Viewed/Week	0.007	0.318	43.8%	68.8%
Total Number of All Discussion Post Types	0.004	0.422	56.3%	62.5%
Week 13 Logins	0.009	0.402	75.0%	62.5%
Week 11 Logins	0.013	0.372	75.0%	62.5%

To answer Question #3, based on the student data provided for this study, which variables are more relevant to others, the results of the correlation analysis and the independent variable that had the most direct relationships with the other independent variables was combined. Although the variables identified as being relevant were not similar between the correlation analysis and the Bayesian network graph, the results indicate that the successful student will be engaged with course material, and will revisit content frequently.

6.4 Discussion

With the projected adoption of learning analytics in the very near future within educational institutions (Johnson, et. al., 2011), the evaluation of prediction variables and their relevancy in identifying students at risk will assist with the continued development of student success models and prediction tools. Although student demographic information and previous academic history or performance data is often included in prediction models based on empirical research, our research showed that variables related to student behavior have often higher relevancy in successfully identifying students at risk. The findings of this study confirmed that student background data may assist in classifying students at risk early in a semester; however student behavioral data, specifically the engagement and interaction of students via discussion board forums are the most relevant variables in successfully identifying students at risk.

There are limitations with this research, specifically related to the size of the data sample and the student success rates. The data was provided from one single institution. Additionally, the course material and subjects from the nine courses used in the study was diverse but from the same program. The courses were only a subset of those available within the program, and even smaller from the total number of courses offered by the institution as a whole.

7. CONCLUSIONS

This chapter illustrates and identifies several key problems faced by learners. In Section 2 of this chapter, a means was described of improving the precision of learning style and WMC identification based on behavior patterns - which are general to any learning management system. Several algorithms were tested which improved the precision of identifying learning style and WMC. The results uncovered that precision is improved by finding

an optimal set of weights for the behavior patterns. Furthermore, it was found that the behaviour patterns are successful as direct inputs into a classification algorithm.

The third section introduced a system that supports learners in collaborative settings, providing them with recommendations and information about how they can learn more effectively as a group. The system focuses on project management and communication aspects and supports individual learners as well as the whole group.

The concept of adaptation was also addressed in Section 4, the notion of adaptive context-aware learning systems was discussed. A framework was proposed that would enable the integration of an adaptive context-aware system with the inference engine of an expert system. The resulting framework would allow for quicker development of such systems with minimal work on the part of the researchers.

Section 5 proposed a solution that would help with the acquisition of scientific competences, which is a key issue in postgraduate programs. The proposed system provides recommendations to beginning researchers, supporting them in doing research.

In Section 6, the problem of high failure and dropout rates for online distance education courses was discussed. In the proposed research, a variety of variables was investigated in terms of their relevance for identifying students who are at risk of failing a course..

REFERENCES

Anagnostopoulos, C. N. E., Hadjiefthymiades, S. (2009). Advanced Inference in Situation-Aware Computing. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans,* 39(5), 1108-1115.

Brindley, J. E., Walti, C., and Blaschke, L. M. (2009). Creating Effective Collaborative Learning Groups in an Online Environment. *The International Review of Research in Open and Distance Learning*, 10(3), 1-19.

Broadway, J. M., & Engle, R. W. (2011). Lapsed attention to elapsed time? Individual differences in working-memory capacity and temporal reproduction. *Acta Psychologica*, 137(1), 115-126.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5-6), 318-331.

Chang, T. W., El-Bishouty, M. M., Graf, S., & Kinshuk (2013). An Approach for Detecting Students' Working Memory Capacity from Their Behavior in Learning Systems. In *13th IEEE International Conference on Advanced Learning Technologies* (ICALT), (pp. 82-86).

Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. Engineering Education, 78(7), 674-681.

Ford, N., & Chen, S. Y. (2001). Matching/mismatching revisited: an empirical study of learning and teaching styles. *British Journal of Educational Technology,* 32(1), 5-22.

Gathercole, S. E., Alloway, T. P. (2008). *Working Memory and Learning: A Practical Guide for Teachers.* London: Sage Press.

Graf, S., Lan, C. H., Liu, T. C., & Kinshuk, (2009). Investigations about the effects and effectiveness of adaptivity for students with different learning styles. In Advanced Learning Technologies, 2009. In *Ninth IEEE International Conference on Advanced Learning Technologies* (ICALT), (pp. 415-419).

Graf, S., Kinshuk, & Liu, T. C. (2009). Supporting teachers in identifying students' learning styles in learning management systems: An automatic student modelling approach. *Educational Technology & Society*, 12(4), 3–14.

Graves, M. F. (1976). Practical Problems of the Beginning Researcher. *Research in the Teaching of English*, 10(1), 51–57 CR.

Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K., (2011). The 2011 Horizon Report. Austin, Texas: The New Media Consortium. Retrieved from http://net.educause.edu/ir/library/pdf/hr2011.pdf.

Koh, M. H., Barbour, M., Hill, J. R. (2010). Strategies for Instructors on How to Improve Online Groupwork. *Journal of Educational Computer Research*, 43(2), 183–205.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence* (IJCAI), (Vol. 14, No. 2, pp. 1137-1145).

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 4, 950-957.

Liu, G. Z., & Hwang, G. J. (2010). A key step to understanding paradigm shifts in e-learning: towards context-aware ubiquitous learning. *British Journal of Educational Technology*, 41(2), E1-E9.

Lokken, F., & Mullins, C. (2014). Trends in e-learning: Tracking the impact of e-learning at community colleges. *Washington, DC: Instructional Technology Council*.

Mazza, R., Dimitrova, V. (2004). Visualising student tracking data to support instructors in web-based distance education. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 154-161). ACM.

McNutt, L., & Brennan, M. (2005). Work in Progress-Learning Styles and elearning, what is the Connection?. In *Frontiers in Education, 2005. FIE'05. Proceedings 35th Annual Conference* (pp. F1H-F1H). IEEE.

Mitchell, T. M. (1997). Artificial neural networks. *Machine learning*, 81-127.

NPA Core Competencies Committee. (2009). *The NPA Postdoctoral Core Competencies*. New York, New York, USA. Retrieved from http://www.nationalpostdoc.org/images/stories/Documents/ToolkitDocuments/Core Competencies - 10.02.13.pdf

Popescu, E. (2010). Adaptation provisioning with respect to learning styles in a Web-based educational system: an experimental study. *Journal of Computer Assisted Learning, 26*(4), 243-257.

Reeves, T. C. (2000). Enhancing the worth of instructional technology research through "design experiments" and other development research strategies. *International Perspectives on Instructional Technology Research for the 21st Century*, 27, 1-15.

Schaeffer, C. E., Konetes, G. D. (2010). Impact of learner engagement on attrition rates and student success in online learning. *International Journal of Instructional Technology and Distance Learning*, 7(5), 3-9.

Schilit, B., Adams, N., Want, R. (1994). Context-Aware Computing Applications. *First Workshop on Mobile Computing Systems and Applications*, WMCSA 1994, 85-90.

Shaukat, S., Siddiquah, A., Abiodullah, M., Akbar, R. A. (2014). Postgraduate Students' Attitudes towards Research. *Bulletin of Education and Research*, *36*(1), 111–122.

Shih, J. L., Chuang, C. W., Cheng, J. J., Hwang, G. J. (2010). The Development Of Mobile Learning System For Local History And Geography Of Taiwan. In *2nd International Asia Conference on Informatics in Control, Automation and Robotics*, 3, 295-298.

Siemens, G. (2010). What are learning analytics. *Retrieved from* http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

Williams, S., and Roberts, T. S. (2002). Computer-Supported Collaborative Learning: Strengths and weaknesses, *Proc. of the Int. Conference on Computers in Education*, 328–331.

Zuber-Skerritt, O. (1987). Helping postgraduate research students learn. *Higher Education*, *16*(1), 75–94. Retrieved from http://doi.org/10.1007/BF00139249

ACKNOWLEDGEMENTS

The authors acknowledge the support of Alberta Innovates Technology Futures (AI-TF), Alberta Innovation and Advanced Education, NSERC, and Athabasca University.

Chapter 5

ENHANCING MATHEMATICAL PROBLEM-SOLVING EXPERIENCES THROUGH LEARNING ANALYTICS

Rébecca Guillot, David Boulanger, Jérémie Seanosky, Vivekanandan Kumar, Kinshuk
School of Computing and Information Systems, Athabasca University

Abstract: Even with all the technological advancements, a majority of students in today's classes still do their mathematics exercises on paper. This approach does not provide teachers with much information about the way students worked on their exercises. Given this, how could teachers understand the problem-solving process undertaken by students? How could they know where in the process students struggled? How could they understand students' competence, confidence and metacognitive abilities? How could the teachers be given real-time information about the challenges faced by students? How could teachers consume this information and in a timely manner translate it into appropriate support and intervention to the students in need? MATHeX is a learning analytics tool that aims to capture as much real-time data about students' work as possible to better understand their study processes and the relation between study processes and learning outcomes. Tracking such data allows for an in-depth analysis of competence, confidence, and targeted metacognitive abilities of students in underlying mathematical concepts as well as a wider view of their study behavior. These data can then be analyzed, transformed and displayed in an interactive dashboard for the consumption of students, teachers, parents, and school administrators. Would learning analytics be able to enhance the role and effectiveness of a human tutor? Would students be able to motivate and regulate themselves better? Would parents and administrators be able to gauge the overall progress of individual learners towards targeted learning outcomes? These are some of the questions addressed in this research under the scope of learning analytics.

Key words: mathematics experiences, learning analytics, data collection

1. INTRODUCTION – WHY IS IT SO HARD TO SUCCEED IN MATH?

Mathematics is a foundational subject in education. "Mathematics is often challenging for students with and without disabilities to master" (Little, 2009). It is a challenging subject to learn for many students because of four main reasons outlined below.

1.1 Building on previous knowledge

Difficulties in mathematics build up faster than any other subject since mathematics constantly challenges students to make statements and assumptions. Boaler (2013) indicates that new synapses are formed in the brain every time a mistake is committed by a student in mathematics and when the student thinks about why something is wrong. The reasoning process in mathematics is extremely intense, and if not well supported, will soon lead to a rapid fall in interest and capacity. "Math, more than most subjects, is based on sequential learning, and if students have missed any previous concepts, then they are unable to understand new ones" (Gregory, n.d.).

Learning mathematics is similar to building a tower. Each stone has to be securely placed for the tower to be unswerving. When a student studies a new mathematical concept, each of its pre-requisite concepts, co-requisite concepts, and other related concepts have to be well understood and reasonably mastered through diversified explanations and exercises. Further, the relevancy of the concepts and the seamless application of the concepts should also be mastered before the student feels at ease about mathematical problem-solving. Joseph's (2009) study "shows that students must possess relevant knowledge and be able to coordinate their use of appropriate skills to solve problems."

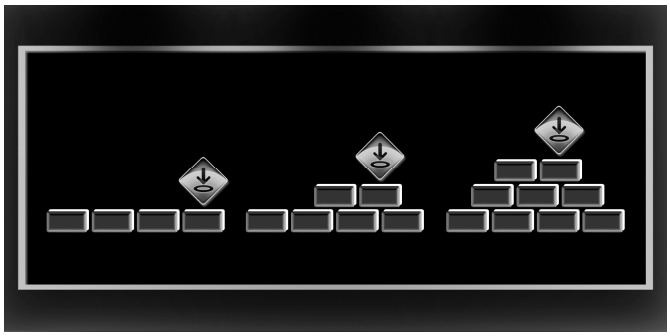


Figure 1- Sequential Learning Tower in Mathematics

1.2 Small procedural errors can lead to a destructive loop related to a loss of confidence

Chinn (2012) states that "many procedures that are taught for mathematics are very unforgiving on faulty memories. Often, even one small error in the application of a procedure is enough to generate failure." Even a small detail could have a substantial impact in mathematics, particularly when students who possess poor conceptual skills continue to solve an exercise using a rather wrong pathway without appropriate feedback. In such an incorrect state, students tend to make more erroneous personalized knowledge construction. Thus, it is important to address such wrong pathways, at real time, before they significantly affect the confidence of the student.

1.3 Confidence and performance are in a symbiotic relationship

Maher (1999) postulates that subject matter understanding emerges when new ideas fit into a larger framework of previously-assembled ideas. A metaphor that reflects this quite well is the notion that one assembles ideas in one's mind much as one assembles a jig-saw puzzle. Each new candidate piece, like each new idea, can be used only if it fits into the aggregate of pieces that have previously been assembled. Similarly, being constantly challenged to make decisions about what is correct and what is not, as a result of the assumptions the student has made while studying, the student may be confronted by some of the solutions and the process that led to the solutions that contradict some of the beliefs, thus affecting confidence. "The implication is that teachers and mathematics educators should focus on the possible difficulties faced by the students as they interact with the mathematical problem and problem solution" (Joseph, 2009). The support from teachers during and after a problem-solving process is crucial because one of the major factors that explains differences in self-confidence between children is their estimated competence (Nunes, Bryant, Sylva, & Barros, 2009). Estimated competence is often misjudged and students "routinely regard mistakes as indicators of their own low ability" (Boaler, 2013).

1.4 Need to support a student in real-time

Mathematical competence building is based on a personalized knowledge construction process. The ability to guide the learner to align, adjust, and consolidate the built competence is known as competence-based learning.

Maher (1999) indicates that a "greater recognition of the many thinking processes that must take place when anyone attempts to deal with a mathematical problem" should be part of the classroom practice. "Teachers, by listening to students, can have close contact with the ideas students are building in their minds. In this way, teachers can try to guide in the construction of those ideas" (Maher, 1999). Contemporary classrooms do not equip teachers to have access to ideas that students build in their minds as they solve problems.

During this mathematical knowledge construction process, the student will be challenged to build a highly personalized understanding of each mathematical problem-solving process. Inevitably, as part of the learning, the student will face facts that contradict the personalized knowledge being built. Maher affirms that "the student who is learning mathematics is supposed to build up a collection of ideas in his or her mind, but, in doing so, encounters fundamental ambiguities" (Maher, 1999). If these contradictions are not dealt with by the student or not detected and addressed by the teacher, they would weaken and erode the student's confidence over time. When these contradictions reach a threshold, a significant portion of confidence in mathematical problem-solving tends to collapse and the student would start to doubt the very construction process as well as the validity of the remaining personalized mathematical knowledge.

In general, only a particular set of points have to be aligned, adjusted, and consolidated in the personalized knowledge to retain the integrity of the entire knowledge structure. Stecker and Fuchs (2000) have shown that "student performance increased when teachers made instructional adjustments based on individualized curriculum-based measurement data." They conclude that "frequent assessment and linked instructional interventions are essential to increasing student mathematics performance."

It has been demonstrated by Parsons that the student performance plays a key role in building self-confidence because "the most important source of self-efficacy was found to be students' past experience of success or failure" and "it is argued that lecturers and support tutors might do more to develop students' confidence" (Parsons, Croft, & Harrison, 2011).

Competence-based learning (Levy & Ramim, 2015; Martinez, Avalos, Lopez, & Palacios, 2015) allows one to measure, predict, and address competences at various levels of granularity. It is important to observe and guide the evolution of competences during the process of their development. That is, as Joseph (2009) points out, students need to problematize their own learning, as and when learning happens.

In general, competence-based learning facilitates continuous and close monitoring of students' study and problem-solving activities to ensure that they learn positively and effectively. In today's classroom, it is impossible to

give extensive individual attention to every student given the number of students in each class as well as the amount of time available during school day. In contemporary instruction, competence can only be assessed when students are assessed. Even if there could be one teacher per student during class, it is near impossible for the teacher to know about the student's homework habits and problem-solving processes. To be a perfect tutor, one would have to be, consistently if not always, available, seeing students' work, answering, and ready to provide useful hints, motivating instructions, and appropriate guidance. Clearly, this is impossible for a human teacher to offer, given contemporary instructional setup! Students are therefore left alone in their learning process and are taught from a distance even in a classroom situation.

This lack of close real-time support is the likely reason many capable students lose interest in mathematics.

2. CHALLENGES FOR TEACHERS

To know what the student knows is a key challenge for mathematics teachers. A more compelling challenge is to receive this piece of knowledge at the right time. An even bigger challenge is to obtain the study context within which the student is trying to acquire the knowledge. This study context relates to contextual elements surrounding the student as he or she is studying. For example, a student who is studying new mathematical concepts while being anxious because of family circumstances or health issues is clearly in a quality-altered study context. Finally, the most critical challenge is to positively engage the student in a pedagogically optimal manner to help gain targeted competences in mathematics. Teaching, explaining, gesticulating, animating, and overwhelming the student with instruction without this positive engagement would only yield a marginal improvement in the student's learning process and competences. In other words, learning has to be fully bidirectional: teachers giving full information to the students and students giving full information to the teachers. The notion of "fullness" indicates that in addition to information on teaching, learning and assessments that teachers and students share, one could also share information on alternative teaching methods for a group of students in the current learning context, learning techniques being followed by a student, social contexts of learning appropriate for a domain, personalized instructional mechanisms, and so on. In general, in current learning contexts, students are not able to give additional information about their learning, in addition to what can be gleaned from assessments, to teachers.

The greatest quality of a mathematics teacher is to have a good sense of students' understanding, to assist in the building and mastering of concepts, and to groom the personalized knowledge and logical thinking process while addressing the gaps in knowledge and process. This is what is referred to as "Listening to students' mathematical thinking" (Suurtamm & Vézina, 2010). How could a teacher possibly know what is going on in the head of each student? By marking the homework? Unfortunately, in contemporary education, students typically go through problem-solving processes and arrive at impasses or solutions to exercises, without sharing the processes that lead to impasses or solutions with teachers. These hidden efforts of the student and the unaccounted challenges the student faces are the root causes of confusion, discouragement, and failure among students.

Researchers have come to the conclusion that "teachers have an important role in guiding students' mathematical development by engaging them in problems, facilitating the sharing of their solutions, observing and listening carefully to their ideas and explanations, and discerning and making explicit the mathematical ideas presented in the solutions [Ball, 1993; Lampert, 2001; NCTM, 1991; NCTM, 2000]. Several research projects (see for example, Cobb, Wood, & Yackel, 1992; Fennema et al., 1996; Franke & Kazemi, 2001; Simon & Schifter, 1991) have found significant benefits when teachers attend to their students' mathematical thinking. The benefits included higher levels of conceptual understanding by students and more positive attitudes held by both teachers and students towards mathematics" (Suurtamm & Vézina, 2010).

Here is a typical example. A student begins his homework of multiplying polynomials. At first, he is puzzled and does not remember even what a polynomial is. After searching the internet and then going through the textbook, he reads an explanation that helps him to start the homework. But the material he has read has also raised more questions in his understanding. Is there always a coefficient in front of a polynomial? Is it possible that a polynomial has no variables? Are addition and subtraction operations included in a polynomial? In spite of these questions, the student begins to work on a polynomial multiplication problem. As he begins, he is still confused about the operation that must be completed first – is it the multiplication or the addition. He gives it a try and solves the exercise by doing the addition first followed by the multiplication operation, but the result seems to make no sense for him. He then tries again with a different sequence, again the result seems odd. He doesn't remember if exponents in likely terms must be added or multiplied when multiplying the terms. Finally, he decides to compute his answer with an online mathematics application on the internet. The online application offers a solution. After writing down the answer, he finds that his current answers are quite different from this solution. He is not sure how else

the solution can be arrived. He stops working on the problem and postpones it for another time. The negative thoughts are already haunting him, making him suspect his mathematical abilities compared to his classmates. The next day, a smiling teacher works out the solution for the class and asks if anyone has any doubts. The student decides not to expose himself to the mockery of his classmates and does not reveal his doubts or the negative thoughts. He convinces himself that he will probably understand the polynomials concept down the road.

This example displays the 'one way' interaction where the student has not been able to give information to his teacher thus preventing this positive engagement needed for success. As outlined in this example, the current approaches to classroom mathematics education do not provide the teacher with critical information that is needed to understand the learning and problem-solving processes of a student. Without knowing it, the teacher is not in a position to engage the student in a deeper learning process.

One can imagine how the scenario would have been different if the teacher had followed the entire problem-solving process of the student during the evening and having chatted in real time with the student in private or even having a personal talk with him in the next morning before the start of the class about the issues he faced. Capture of the entire problem-solving process could be a gold mine for a teacher who wants to 'hear' about the knowledge of a student. These 'captured' exercises could potentially identify the weakness in the student's understanding and application of mathematical concepts.

The sequence of solution steps submitted by the student and the final answer of the student only reveal a small portion of the student's skills and knowledge gaps. Teachers need to have access to the 'captured' problem-solving habits and challenges of the students. Students' habits and challenges offer a way for the teacher to pinpoint concepts that need to be reinforced and mastered before going any further.

Mathematical weaknesses have the tendency to be covered up if the whole process of problem-solving is not 'visible' or 'audible'. This is why it is necessary for an instructor to have a deeper and timely understanding of students' progress in order to spot weaknesses as and when they occur.

The next biggest challenge for a teacher is to properly address the weaknesses related to mathematical concepts among his students at the right moment with appropriate evidences as to the necessity of such an address. The teacher needs to provide timely and personalized support according to each students' individualized needs and difficulties.

With the current structure of the education system, it is extremely difficult for teachers to respond appropriately to every difficulty encountered by

students. More often than not, teachers are unaware of the existing difficulties among their students.

3. CHALLENGES FOR STUDENTS

As mentioned before, mathematics constantly challenges the mind. While studying mathematics, each student is filled with questions, assumptions, hypothesis, conclusions, and statements. Each student also exerts a certain level of understanding and the ability to apply that understanding in problem-solving situations. Teachers need to ensure that students' comprehension is understood, grounded, and mastered. To optimize the teacher-student interaction, students need to have access to an open channel to share their questions, concerns, confusion, and successes as they study or work their math exercises. There are many crossroads during the process of problem-solving and students are expected to recollect and summarize past study experiences into the current problem solving context. The biggest challenge for students is to undergo the learning process with few inputs from their teachers and with few outlets to communicate the details that make them perplex and alter the validity of their answers.

It is very hard and discouraging for students to work on their own with no access to someone to whom they can ask for help or express difficulties, but rather having to rely on their own judgement and assumptions. Moreover, students who face difficulties in their work are not able to pinpoint the exact problem they faced. This leads to deterioration of their motivation and capability to share their difficulties with someone else. In general, students facing difficulty end up with a general statement such as 'I do not understand my math'.

Even students capable of logical thinking and mathematical reasoning may have under-expressed or suppressed misconceptions. Students may not have the capacity to recognize weaknesses and strengths. Students may emphasize more on their weaknesses than strengths, thus resulting in two significantly different viewpoints, a students' viewpoint and a teacher's viewpoint, as shown in Figure 2. While the teacher is able to discern the few concepts that need to be reinforced, the student often exaggerates the extent of a few misunderstandings into a generalized situation, and thus, based on a select set of bad results, concludes being incompetent in mathematical problem solving.



Figure 2 – A Student and Teacher's Different Viewpoints

Therefore, it is important that students express even little parts of problem-solving that confuse them as they work through a mathematical exercise. It would be a welcome change if students' problem-solving episodes along with specific difficulties that they faced are captured in an automated fashion, compiled, and presented to instructors through an application.

Such an advanced application, called ALEKS, is a Web-based system, working around "an artificial intelligence engine that assesses each student individually and continuously," "mapping the details of each student's knowledge." ALEKS is monitoring the learning process of each student and is able to provide the student with a "selection of only the topics" he is ready to learn at this exact moment. ALEKS also records successes and failures to guide the student for an optimal learning path through "one-on-one instruction, 24/7" (McGraw-Hill, 2015).

While ALEKS focusses on knowledge aspects, it does not provide an analysis of both the level of competence and the level of confidence of students which is crucial in learning mathematics.

4. MATHeX SOLUTION

MATHeX is a learning analytics tool that aims to study students' conceptual, problem-solving, and metacognitive behavior as they work through mathematical exercises. It is a companion software that is always there whenever students study mathematics and solve problems, 'observing' and 'listening' to their challenges, guiding them even in individual steps, giving them instructions, providing encouragement, and displaying their progress. The goal of MATHeX is positive engagement with students. The tool also focuses on finding ways to strengthen the study habits of students, creating

opportunities for motivation and continuous informal assessment of their progress.

MATHeX features

MATHeX features can be summarized by these two key points: capturing data and analyzing data. These two points are explained in more detail below.

4.1 Capturing data

MATHeX aims to collect as much data as possible from the observations of students' mathematics related work. The tracking of data by MATHeX supplements the data inherently collected by the teacher, either in the form of listening to a student or classroom observations or assessment results. The collection of student's data is crucial in engaging the students in their learning process thus enabling and enhancing bidirectional communication between the teacher and the students.

The current version of MATHeX captures the following datasets:

4.1.1 Timestamps

The time dataset provides information on each study session (including reading and preparing to solve mathematical problems), the amount of time spent on each step of a problem-solving process, and the overall duration in each problem.

MATHeX also tracks the time of inactivity within a problem-solving process. After a preset time of inactivity, the software will invite the student to indicate the reason for inactivity from a list of options. The options include a time of absence from the study session, a time of reflection about the mathematical problem, a time of 'silence' created by confusion, and a time of being stuck needing help. This feature will also help estimate the attention span and concentration skills of students.

4.1.2 Pointer, mouse, menu, and keyboard activities

As the student works through exercises or assignment problems on a computer, including mobile devices, MATHeX tracks pointer, mouse, menu, and keyboard status of the devices. For example, mouse click statii during a problem-solving activity is captured and associated with the corresponding problem the student is attempting to solve. These datasets include the erasing of an answer, the reviewing of the problem statement, the changes made

within a solution step, the help-seeking behaviour, and so on. This capacity to capture all writing, typing, and graphical activities on any computer with the standard internet browsing capacity is key to 'listening' to the student's mind during problem-solving processes. MATHeX aims to discover patterns in the solving process of each student, to have a better understanding of weaknesses, hesitations or skill achievements. In capturing the expressed (written, moused, typed, or selected) steps of a given problem, MATHeX is able to detect the confidence and the competence of the students in solving a particular problem type.

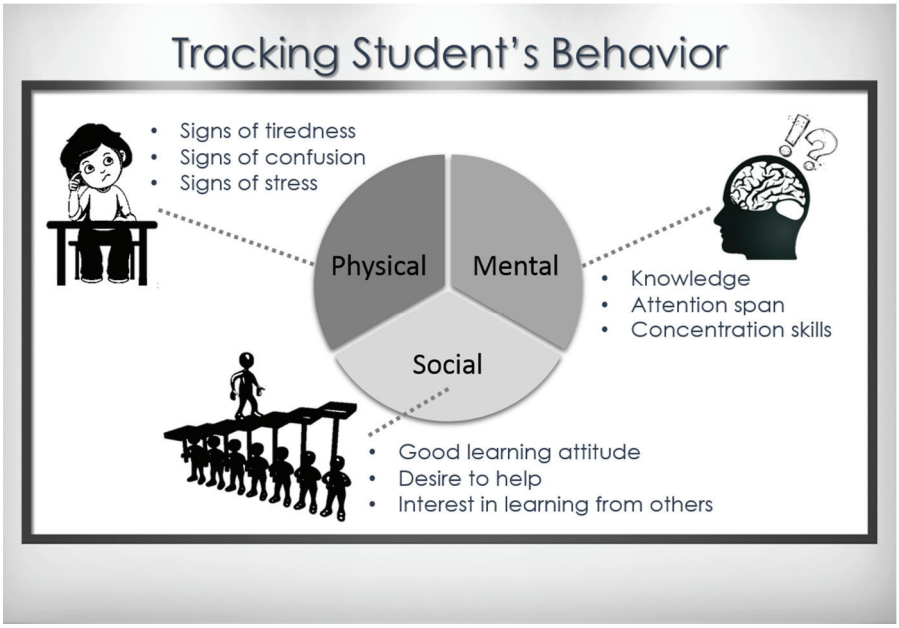


Figure 3 – MATHeX Tutoring Environment

MATHeX offers multiple learning environments such as individual, tutoring, social interactive, and 3D virtual learning spaces. Activities within each of these spaces are sensed and used to recognize problem-solving patterns.

4.1.3 Contextual information

The context of mathematical problem-solving experiences in MATHeX includes information such as the types of problems attended by the student, topic space covered by each problem, the study patterns of the student, the goals of the teacher, the help-seeking behaviour of the student, and the student's interactions with the problem-solving environment. In future work,

MATHeX aims to capture visual data on students' eye movement and facial expression, as further validation of signs of confusion, stress or tiredness. Information that are not available through mouse and keyboard such as the heartrate and other physiological data can be attached to the context from wearable devices.

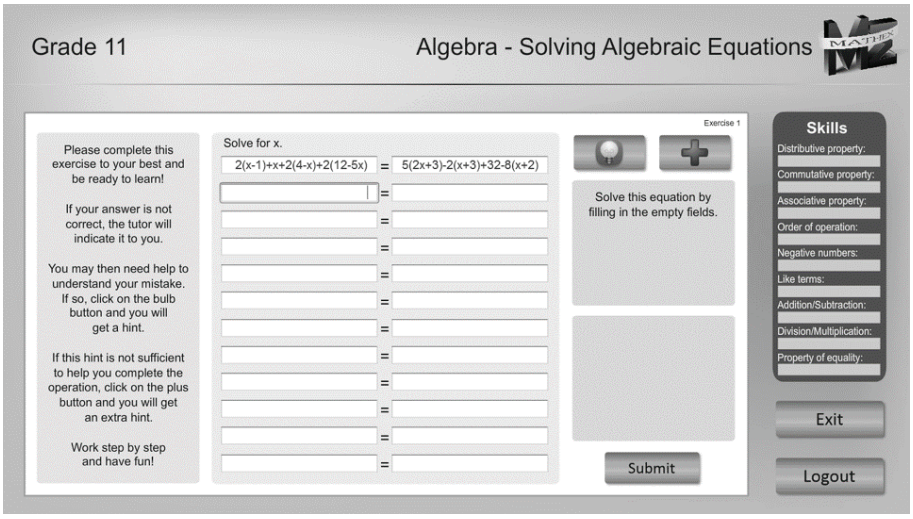


Figure 4 – Student's Behavior Data

MATHeX also aims to capture the social behaviour of students while engaged in problem-solving activities. Metacognitive traits such as self-regulation and co-regulation can be observed from within MATHeX with assistance from built-in tools such as Self- and Co-Regulated Learning (Zheng, 2015).

4.2 Analyzing data

Captured data are subjected to continuous analysis in MATHeX involving data transformation, data visualization, and regulation.

4.2.1 Transforming data

MATHeX uses an analysis engine named SCALE (Boulanger, 2015) that transforms raw data in meaningful data. The analysis results in an assessment of students' competence and confidence based on the observed mathematical problem-solving experiences.

The analysis engine (SCALE) receives structured data in a MathML content format and will then analyze the validity of the answer and the overall problem-solving process of the student.

MATHeX aims to measure weaknesses identified in a student's understanding in terms of targeted skills. Observations on the speed and accuracy of the problem-solving process, in conjunction with the correctness of the solution, can be mapped to individual problem-solving skills. Such an individualized observational process guides the student through needed supplemental exercises and activities. The growth of individual skills can be transformed into competences using a competence framework such as the Bloom's Taxonomy (Maker & Nielson, 1995; Highley & Edlin, 2009).

In the tutoring environment, a competence-based analysis will assist in the offering of real-time feedback to the student.

4.2.2 Visualizing data

Data can be valuable if they are visually accessible and meaningful. MATHeX offers displays of pertinent information for students, teachers, parents, and school administrators in an interactive dashboard. This visualization is key to engage students and motivate them to reflect on their performances throughout the learning process. MATHeX dashboard also provides visuals on the competences of the student. MATHeX's teacher dashboard enables an overall picture of the entire class. The dashboard will also enable teachers to communicate with students at real-time to offer feedback on specific competences and struggles experienced by students.

The dashboard will help parents to see ranking of a student in comparison with the rest of the class.

The student's view of the dashboard provides information that associates a student's problem-solving performances with the expected outcomes.

4.2.3 Self- and co-regulation

MATHeX includes a self-regulation feature that allows students to create their own initiatives, in a tool called SCRL (Zheng, 2015).

Allowing students to regulate their own study behaviour engages students in their activities as no teachers, as motivated as they can be, would be able to do. MATHeX encourages student's self-regulation knowing that "undoubtedly, all learners are responsive to some degree during instruction; however, students who display initiative, intrinsic motivation and personal responsibility achieve particular academic success" (Zimmerman, 1990).

Students will have the opportunity to motivate and regulate themselves by interacting with their competence levels in the dashboard and by creating their new initiatives, setting new goals, picking out strategies to achieve these goals, and actually achieving these goals. Students will be accompanied by

MATHeX till the completion of their initiatives and will be guided for better success.

MATHeX also aims to provide a co-regulation aspect that would encourage peers to help each other and assist one another with respect to individual initiatives. This aspect will also create new datasets as to the social engagement of students with their classmates. DiDonato's (2006) study reveals that when students help each other, they learn tremendously and they are encouraged to develop socially. Similar results have been noted in the peer tutoring research (Kumar, 1996). Consistent with the old adage: 'to teach is to learn twice,' research on peer tutoring has found increases in both the tutee's and tutor's academic and social development as a result of peer tutoring interventions (DiDonato, 2006). "In these cases, there are a number of academic and social benefits to all group members as a result of participating in co-regulatory processes." That is, co-regulation may have benefits for both the person doing the regulating and other group members to whom the action is directed, and this may lead to increases or refinements in both students' SRL" (DiDonato, 2006).

Moreover, Perger (2013) also came to the conclusion that "working with others was another practice recognized by both adult experts and students as important when learning mathematics. Teachers need to develop learning environments and practices that encourage students to work in groups. The teacher works as a co-ordinator providing guidance and support both in mathematics content learning and in developing skills that enable students to work together. This ability to work together has been recognised as a skill students need to be taught" (Perger, 2013).

Mathematics is a complex subject and at the same time can be explained in simple terms. Experiencing the same concept in different contexts might enlighten a confused student. With this perspective in mind, Boaler (2013) expresses how "encouragement of a growth mindset culture will require schools to move to grouping practices that do not label or send negative messages to students, and teaching approaches that value the thinking, struggles and varied learning pathways of all students."

5. CONCLUSION

Many online applications allow students to learn mathematical concepts and train themselves. Among those is ALEKS, which focusses on the knowledge aspects (McGraw-Hill, 2015). However, this research aims at an analytics tool that would provide more than just an application to learn and be trained. Such an analytics tool allows students to get precise feedback about what they master and what they don't, to be guided through specific exercises that

respond to their individual needs, and to enable teachers to know exactly about the problem-solving process that the students have experienced.

MATHeX aims to provide analysis of both the level of competence and the level of confidence of students. Further, MATHeX looks to capture study habits, social contexts, and inclination of students. In MATHeX, graphs in the dashboard are interactive and allows the students to create their own initiatives through the self- and co-regulated embedded learning system. Moreover, MATHeX enhances mathematics study experiences by allowing students to work any types of problems and capturing their work to analyze both strengths and weaknesses during the solving process. In addition to ALEKS' features, MATHeX helps the teachers understand the problem-solving process undertaken by the students, and know where the students struggled, and what are their competence, confidence, and metacognitive abilities. MATHeX is a tool intended to 'listen' to the students' mind during problem-solving processes thus providing this open channel that will allow the teacher to understand what the students would like to say and engage them positively. In capturing the expressed (written, moused, typed, or selected) steps of a given problem, MATHeX is able to detect the confidence and the competence of the students in solving a particular problem type.

It is important to note that the observed and inferred data from students' interactions can only be shared with others, including teachers, parents, and administrators, with the expressed consent of the student. The data is inherently owned by the student and the student's permission is explicitly sought to use the data for analytics and share the data with others.

Building strong mathematics skills needs tutor involvement, passion, care, time, and devotion. The key interest of a mathematics teacher is to know how the student has solved his problem, what difficulties he faced, how much time he took to be successful, and much more. In today's classroom, it is impossible to give this individual attention to every student. MATHeX wants to bring solutions by being a learning analytics tool that accompanies the teacher and the student, captures the student's overall activities, identifies weaknesses through analysis, guides the student and provides instruction, gives feedback, and displays visually all needed information in a dashboard. The purpose of this study is to know if learning analytics would enhance the mathematics experiences of the student and to discover the role of learning analytics in mathematics education. It is also the goal of this research to learn about the best approach in classroom between the paper and the computer. These approaches will be evaluated using the results of the students and their feelings towards it.

REFERENCES

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *The Elementary School Journal*, 93(4), 373–397.

Boaler, J. (2013). Ability and Mathematics: the mindset revolution that is reshaping education. *FORUM, 55*(1).

Boulanger, D. S. (2015). SCALE: A Competence Analytics Framework. *International Conference on Smart Learning Environments.*

Chinn, S. (2012). Beliefs, Anxiety, and Avoiding Failure in Mathematics. *Child Development Research*.

Cobb, P., Wood, T., & Yackel, T. (1992). Interaction and learning in mathematics classroom situations. *Educational Studies in Mathematics, 23*(1), 99-122.

DiDonato, N. C. (2006). Effective self- and co-regulation in collaborative learning groups: An analysis of how students regulate problem solving of authentic interdisciplinary tasks. *Instructional Science*.

Fennema, E., Carpenter, T.P., Franke, M.L., Levi. L.,Jacobs, V., & Empson, S. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research inMathematics Education, 27*, 403-434.

Franke, M.L., & Kazemi, E. (2001). Learning to teach mathematics: Developing a focus on students' mathematical thinking. *Theory into Practice, 40*, 102-109.

Gregory, L. (n.d.). Yes, Math is harder. Yes, you can pass! *University of Texas at San Antonio*.

Highley, T., & Edlin, A. E. (2009). Discrete Mathematics Assessment Using Learning Objectives Based on Bloom's Taxonomy. *Frontiers in Education Conference, 2009. FIE '09. 39th IEEE*, 1-6.

Joseph, K. K. (2009). Secondary 2 Students' Difficulties in Solving Non-Routine Problems. *International Journal for Mathematics Teaching and Learning*.

Kumar, V.S. (1996). Computer-supported collaborative learning: Issues for research. In proceedings of the 8th Annual Graduate Symposium on Computer Science, University of Saskatchewan, Canada, (pp. n/a).

Lampert, M. (2001). Teaching problems and the problems of teaching. *Journal of Mathematics Teacher Education*, 5(2), 187-199.

Levy, Y., & Ramim, M. M. (2015). The Effect of Competence-Based Simulations on Management Skills Enhancements in E-Learning Courses. *Proceedings of 2015 Chais Conference for the Study of Innovation and Learning Technologies*, (pp. 34E-41E). Raanana, Israel.

Little, M. E. (2009). Teaching Mathematics: Issues and solutions. *TEACHING Exceptional Children Plus, 6*(1).

Maher, C. A. (1999). Mathematical Thinking and Learning: A perspective on the Work of Robert B. Davis. 85-91.

Maker, C. J., & Nielson, A. B. (1995). *Teaching Models in Education of the Gifted.* PRO-ED inc.

Martinez, M. S., Avalos, M. G., Lopez, G. d., & Palacios, G. E. (2015). Curricular Reform: Changing from Disciplined-Based. *International Journal of Education, 7*(1).

McGraw-Hill, E. (2015). *About Aleks*. Retrieved from aleks: http://www.aleks.com/about_aleks

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Professional standards for teaching mathematics*. Reston, VA: Author.

Nunes, T., Bryant, P., Sylva, K., & Barros, R. (2009). *Development of Maths Capabilities and Confidence in Primary School.* University of Oxford.

Parsons, S., Croft, T., & Harrison, M. (2011). Engineering students' self-confidence in mathematics mapped onto Bandura's self-efficacy. *Engineering Education*, 52-61.

Perger, P. (2013). What They Say, What They Do Understanding Student's Perceptions. *International Journal for Mathematics Teaching & Learning*.

Simon, M., & Schifter, D. (1991). Towards a constructivist perspective: An intervention study of mathematics teacher development. *Educational Studies in Mathematics, 22*, 309-331.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: *The importance of individual progress monitoring*. *Learning Disability Research and Practice*, 15, 128-134.

Suurtamm, C., & Vézina, N. (2010). Transforming pedagogical practice in mathematics: Moving from telling to listening. *International Journal for Mathematics Teaching and Learning*.

Zheng, L. E.-B. (2015). A Framework to Automatically Analyze Regulation. In V. K. G. Chen, *Emerging Issues in Smart Learning* (pp. 23-30). Berlin: Springer Berlin Heidelberg.

Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, 3-17.

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

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.

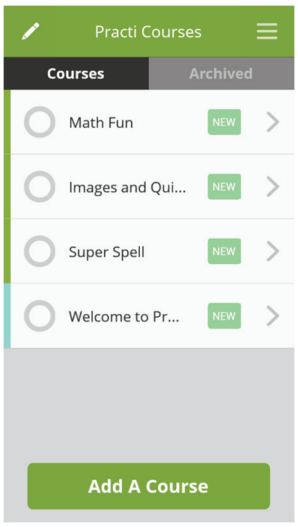
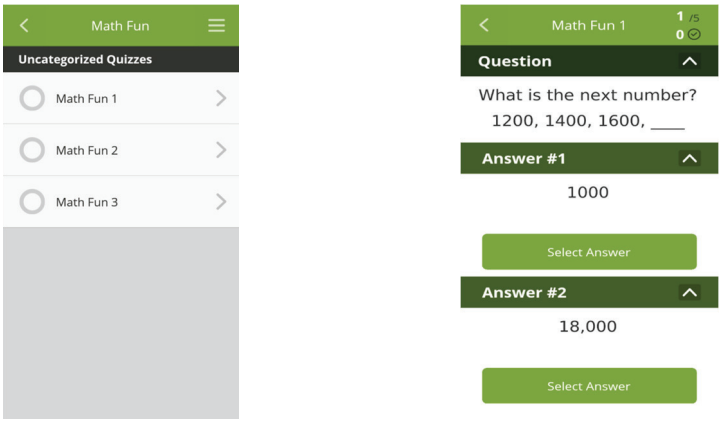


Figure 1. Student has four courses added to his or her *Play*.

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.



(a) (b)

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.

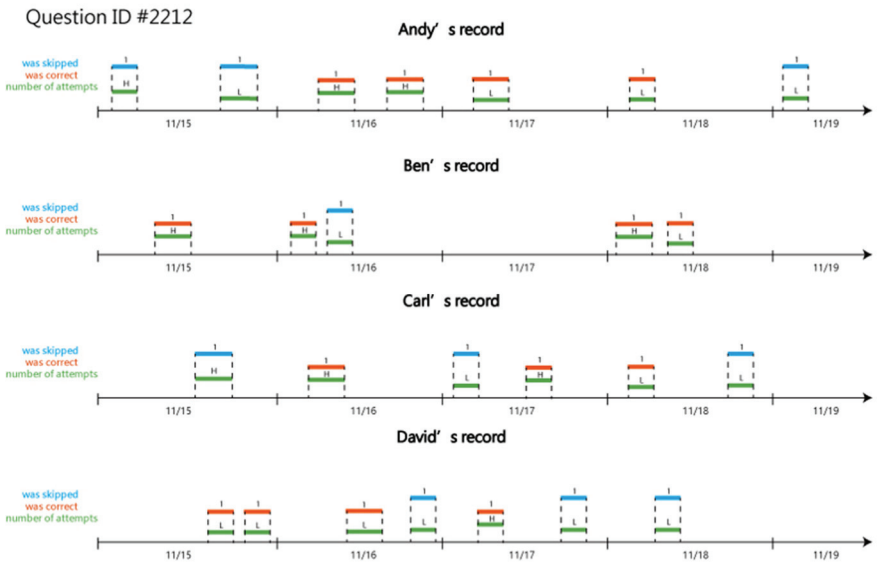


Figure 3. Four students' behaviours of solving question #2212.

Correct or skipped (Correct = 4 ,Skipped = — 4)
Attempts state (H = -2, L = 2)
Time state (H = -1, L = 1)

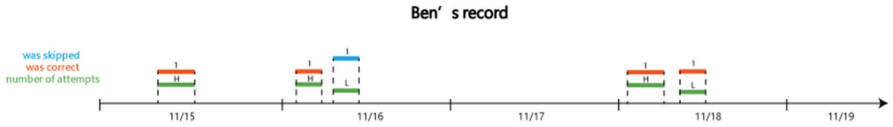
Patterns		Difficulty	Equation	Weight	Symbol
Correct and attempts (L) with time (L)		Easy	=4+2+1	7	A
Correct and attempts (L) with time (H)		Easy	=4+2-1	5	B
Correct and attempts (H) with time (L)		Easy	=4-2+1	3	C
Correct and attempts (H) with time (H)		Normal	=4-2-1	1	D
Skipped and attempts (L) with time (L)		Normal	=-4+2+1	-1	E
Skipped and attempts (L) with time (H)		Hard	=-4+2-1	-3	F
Skipped and attempts (H) with time (L)		Hard	=-4-2+1	-5	G
Skipped and attempts (H) with time (H)		Hard	=-4-2-1	-7	H

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.



Question ID	User ID	Pattern sequence
2212	Ben	D,C,E,D,A

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).

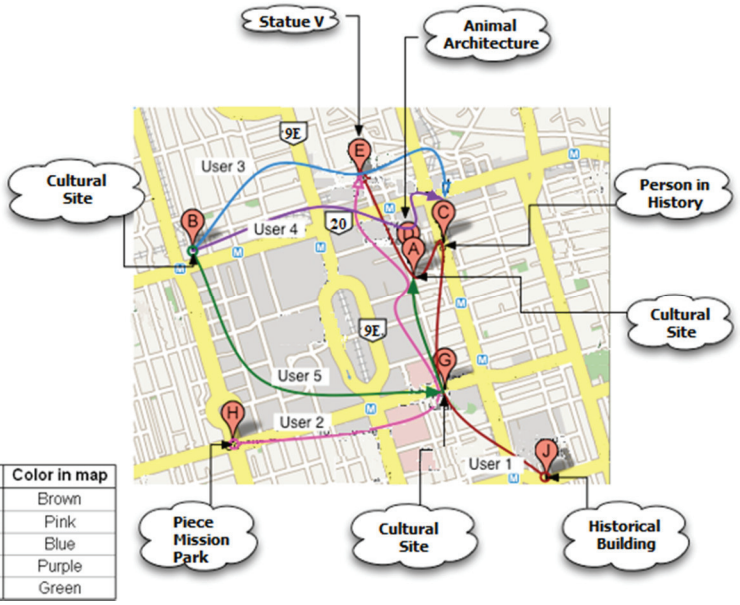
The abovementioned data analytics method can also be adopted to analyse the usage data (i.e., feature usage, actions taken and browsing behaviours) that a system may record and store. The retrieved frequent pattern from the dataset of sequences can be treated as most common behaviours that users may 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 <J G C A E> 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.



User Id	Routes	Color in map
1	<J G C A E>	Brown
2	<H G A D E>	Pink
3	<B E A C>	Blue
4	<B D A C>	Purple
5	<B G A>	Green

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).

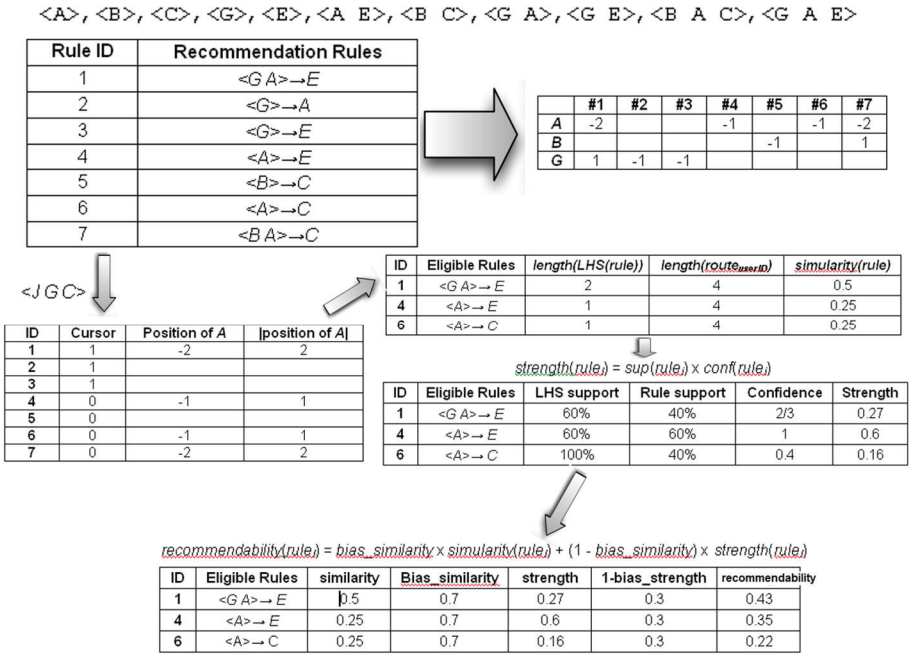
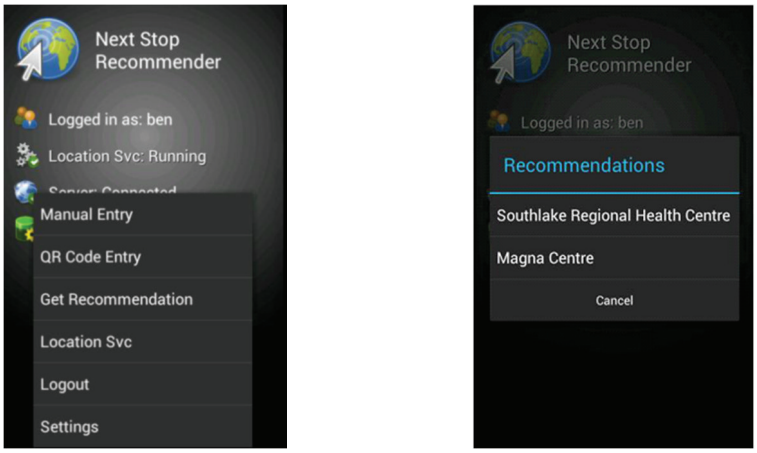


Figure 8. The workflow of finding eligible rules for making recommendations.

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 coodinates 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.



(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.

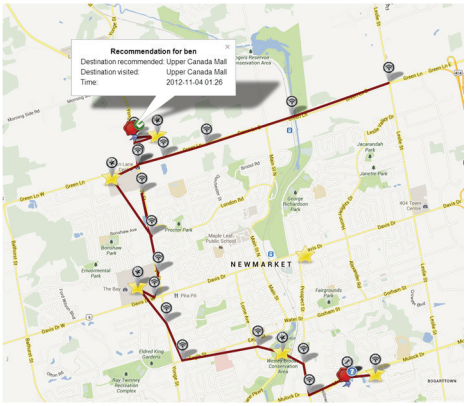


Figure 10. A student took system' recommendation and paid a visit to the place.

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 "AAABBBCCCABBCCCC" 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 JJJJJJJJAAAAAALAAAAA as example, the method can find all possible words with regular expressions' help:

JJJJJJJAAAAJAAAAAALAAAAA

→ (JJJJJJ)(JAAAA)(JAAAA)(AA)(L)(AAAAA)

→ (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 instance, the second generated sequence JJJJJJJJAAAAAALAAAAA has 0.9818717 cosine similarity with the reference sequence JJJJJJJAAAAJAAAAAALAAAAA, higher than the other two generated 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 location-based 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-math-education-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-oecd-finds-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.

Chapter 7

ENHANCING MUSIC PROWESS THROUGH ANALYTICS

Claudia Guillot, Rébecca Guillot, Vivekanandan Kumar, Kinshuk
School of Computing and Information Systems, Athabasca University

Abstract: Analytics is the study of context-aggregate and context-precise insights. These insights are derived from observed user experiences represented in computational scaffolds. The scaffolds allow both humans and software agents to continually validate and characterize the insights. In the setting of learning analytics, these insights trace, employ, analyze, discover, and foster learning traces toward conceptually-optimal, emotively-engaging, cognitively-efficient, and creativity-inspired learning experiences where a learning trace comprises of an instantiated network of models that lead to a measurable chunk of learning. While learning analytics platforms in general focus on the tracking and reporting of broad and generic learning activity events, new e-learning technologies as introduced in the setting of this research have enabled to apply analytics not only in learning in general but also in specific learning domains such as mathematics, English writing, and programming. Since music as a domain requires a high level of monitoring to guarantee learners' success, learning analytics will play a key role in helping music students attain optimal accomplishments. This chapter on music analytics discusses ways to enable teachers to listen whenever and wherever students practice instrumental or vocal music, to identify challenges encountered by students, to track the development of students' skills in solving music theory exercises, and to understand the causes for their successes and failures. This chapter introduces a music analytics tool called MUSIX to analyze the performance of students in terms of proficiency and confidence in music learning objectives and to promote self-regulatory traits to help them manage their own learning processes.

Key words: music analytics – e-learning – teaching approach – audio software – human-computer interaction

1. INTRODUCTION

Music is an art consisting of a suite of unique skills such as vocal singing, playing an instrument, understanding musical notation, and creating a musical score. Musical skills can be nurtured with sustained, adaptive, and goal-oriented instruction and practice, to meld into musical competences. Music teachers help shape the development of individual skills before the skills are honed as competences of musicians.

Traditional music teaching, however, caters only to a limited understanding of the music learning processes and learning challenges faced by students. Wise, Greenwood, & Davis (2011) state that "many secondary school music teachers are products of the Western classical tradition, which is based largely on the conservatoire and the associated skills and traditions that this brings with it" (p.121). That is, music teachers have limited opportunities and datasets to determine the needs of students in terms of individual skills. This is further compounded by the lack of information on the depth at which skills of individual students need to be nurtured.

One may contend that teachers could target instruction at a particular depth prescribed by taxonomies such as Bloom's (Rodrigues & Dos Santos, 2013). But, practicality of teaching indicates that gauging the depth at which students are toiling is an extremely difficult proposition. Such information typically arrives as outcomes of an assessment activity. Teachers also employ formative techniques such as in-class performance observations and group work to instruct students who lag behind targeted skill level.

This is where learning analytics plays a central role, in providing teachers with measurable, continuous, and detailed information concerning the development of individual skills in each student, from both formal and informal study sessions, including in-class, in-the-lab, and at-home sessions. Further, teachers can be made aware of the learning process adopted by students, including the challenges students encounter during each study session. Learning analytics system could also correlate study habits of students with the assessment outcomes, and prescribe optimal study pathways for individual students and advocate compatible pedagogies appropriate for each student.

A learning analytics tool for music, called MUSIX (Guillot, 2015), has been designed and is being developed to address these key challenges. This chapter offers a review of the traditional music teaching approaches and highlights the complementary music teaching approaches using MUSIX's analytics.

2. TRADITIONAL MUSIC TEACHING

While learning music is an immensely rewarding experience, it can also be at times discouraging. Some of the causes of such discouragement are discussed below.

2.1 Lessons to large groups

In many music schools, primarily, lessons are offered to groups of students. Playing music in group or singing with ensemble is a great asset in the training of a musician. However, the group approach does not bode well to detecting students' difficulties or determining the appropriate level of assistance to be provided to students in their daily routine of learning music since every student is unique with different learning capacities and preferences. Typically, students can be identified in one of these three categories: talented, average, or struggling.

Average students are not penalized much in group study or practice since they possess adequate study skills to achieve learning outcomes set forth by the teacher. But, difficulties experienced by the struggling students often go unnoticed. They would require more time and additional instructions to help them achieve the learning outcomes, but they tend to keep silent and not seek teacher's help, thus choosing to remain unnoticed and are inevitably left behind. It is clear that this phenomenon generates discouragement and it may create major gaps in the learning processes of students, even dissuade them to pursue further musical studies.

Talented students tend to accommodate instruction as provided and continue their levels of performance. They, however, can be slowed down in their learning process by the rest of the group. In such cases, this context tends to diminish their interest and attention which is also damaging to their possible future career as musicians.

Essentially, the group approach does not suit all students. However, individualized training and personalized instruction in music education should not diminish the importance of group study. Group study should be considered a significant part of the general music education.

2.2 Lessons with fixed time intervals

Another difficulty with the traditional music teaching approach is the system of lessons offered at fixed time intervals. It is mostly used in the case of individual lessons. Usually, instrumental lesson or voice training lesson are given once a week. Individual teaching is a great way (if not the best) to address the specific needs of a student; it can be advantageous because of the

one-to-one interactions between the student and the teacher. However, having only one hour a week to observe practice habits and growth of skills offers minimal information to the teacher thus leading to non-optimal instruction. This exposes a main concern about this approach. The teacher can assess and guide the student within a study session, but there is no monitoring of the student's study processes and achievements outside of the study session. When left alone to practice, the student can face challenges and be confused about the ways to address them. Maintaining a positive learning experience is rather difficult when the students try to resolve these challenges on their own (often in a wrong way) or ignore them.

2.3 Other disadvantages of traditional teaching

While the world is making huge steps in introducing technology-enhanced learning in many subject areas, music remains one of the few that are still taught in the old-fashioned way. Bauer (2014) describes that "while there is evidence that new, technology-based approaches may be increasing in some areas of music education [Williams, 2012], it appears many music educators are not actively utilizing technology in a manner that could potentially facilitate and enhance musical experiences for students."

As discussed above, traditional teaching could hamper the learning efficiency of students since the teacher is not aware of the difficulties faced by the students while they study music. That is, the teacher is only aware of the progress made by the students when they are assessed. Then, and only then, a small window is opened to the teacher about the learning process of his students. Beatty (2015) states that "the assessment and evaluation of students' achievement in music education remains central for the music teacher in determining the knowledge and skills that students have learned."

Traceable study efforts of students are only observed during individual instructional sessions, which represent only a small percentage of the whole music learning process of the student. The details that ought to be known by the teacher during practice time at home or with friends or anywhere outside the classroom are critically valuable and yet they remain unavailable to the teacher.

It is also important to note that some music training software already exist but they do not capture learning traces. Students do use technology to train themselves with whatever means available to them, including the use of a computer in learning music, but such efforts are never brought to the attention of the teacher.

3. MUSIX TEACHING APPROACH

MUSIX is a learning analytics system for the domain of music. It allows students to learn music at their own pace, and explicitly capture their growth in music competence and music confidence. Bauer (2014) aptly expresses that "the art and science of teaching by skilled educators involves making 'real time' instructional decisions, adjusting on the fly." Adhering to this philosophy, MUSIX, at present, only offers individualized approach to detect students' difficulties and to provide custom solution to these difficulties.

Motivation is one of the key aspects one should consider when exploring a new approach to teach music. Learning music should be facilitated in an atmosphere of enthusiasm and positive engagement. Such an environment should also allow students to express their concerns confidently and openly at a time and place of their choice. MUSIX provides multiple avenues of expression and learning for students and also provides the means to measure the statii of individual musical skills, the contribution of individual skills to competences, the confidence with which musical competences have been expressed, and the ability of students to regulate their own learning habits.

3.1 Capturing data on a continuous basis

MUSIX's goal is to track students' activities through instructional sessions, computer-based exercises, games, quizzes, social network interactions — all created within a learning management system —, and practice time. MUSIX aims to capture as much data as possible based on these activities and process them to recognize study patterns, to identify expressed skills, to predict the growth of multiple skills into competences, and to measure self- and co-regulation efforts.

Tracking of a variety of study related data is a main function of MUSIX since it is these datasets that will provide information that are normally not available in traditional approaches.

MUSIX aims to continually update the profiles of students as and when the datasets become available. The profiles could predict, on a daily basis, students who are lagging behind the rest of the class. The outcomes of such predictive models could be made available to teachers and also to individual students. Students, with assistance from teachers, can engage in self-regulatory and co-regulatory initiatives (see Section 3.6 below) to catch up with the rest of the class, thus avoiding being left behind with the knowledge that they are not alone in their learning process.

MUSIX is geared to collect activity-specific datasets. Presently, MUSIX collects data from the following three activities: study of music theory, vocal training, and the playing of an instrument.

3.1.1 Study of music theory

Many concepts of music theory are often misunderstood by students and a number of them consider theory as an optional area of study. But since the study of music theory "mainly deals with the language and notion of music where it is composed and interpreted" (Aldalalah, 2010), it has been proved that "these music concepts have an important role in establishing the necessary knowledge for interpreting the development stages in music and the mode in which the notation is utilized in various situations." (Aldalalah, 2010) MUSIX offers instructional lessons on music theory at different levels of mastery. The progress of students in theory lessons are continually tracked. Using such learning traces, progress of students are frequently updated. MUSIX collects a wide range of datasets including time-oriented datasets (e.g., time to answer question in a quiz, the type of hints requested, the amount of time spent reading a page, and the number of initiatives taken by the student). MUSIX allows tracing data on any activity of learner engagement. For example, if a game-based approach is used to introduce music theory, then MUSIX will enable the tracking of game-related activities of the student.

Music theory lessons are created and implemented in an online learning management system, allowing students to have flexible access to the content and study activities.

3.1.2 Vocal training

Vocal training in MUSIX presents an interesting challenge because not only one needs to capture raw audio recordings of students' practice but also needs to automatically interpret the recordings in terms of skills and the degree to which the skills have been exhibited. MUSIX offers two types of vocal training, first the sight-singing and second, the vocal techniques. These two are closely related but yet, they need to be trained separately.

In order to measure sight-singing performance of a student, MUSIX will first present a melody on the student's computer screen. It will also offer the student with an audio file that plays the note range along with other options such as "listen to the first note" and "listen to the beat measure" and so on. Subsequently, the student will have the opportunity to start sight-singing the melody and record the same. The student is welcome to record multiple times and each recording will be analyzed by MUSIX for accuracy of notes, correctness of rhythms, and appropriateness of pitch.

Under vocal techniques, MUSIX offers training on note accuracy, with a particular focus in the position of many body parts. Singing techniques are improved with good breathing, position of the mouth, appropriate muscle tension, and more. Since the whole body is transformed as an instrument,

physiology datasets corresponding to specific body parts will be sensed and recorded. An experienced teacher will quickly detect a bad technique by looking or by sensing the throat muscles as the student sings. In addition to such physiology datasets, MUSIX will also record information on students' breathing, delays before breathing, the tuning of each note, the duration of the notes, and other similar data.

3.1.3 Instrument playing

Teachers opt to be by the side of their students as they practice their instruments. Discovering how a student practices is a mystery to most music teachers. In unfolding this mystery, MUSIX will capture data such as the timing of a student's instrument practice, the breaks in between practices, the sequence of the practice, the manner with which the student handled errors, the mastery of concepts learned in a lesson, and the number of notes played correctly. McPherson and Renwick (2000) pointed out rightly that "how children plan and manage their time has important implications for how efficient their practice will be".

MUSIX records all possible study activities with respect to time, accuracy of notes, accuracy of rhythms, and other such valuable data. Music related data uses a MIDI connection between a digital instrument and MUSIX in order to transfer converted data. For instance, the digital piano may send many datasets on the notes played, speed at which melodies are played, the articulations of notes, and the rhythms. Other instruments such as a silent violin or an electric guitar will also be capable of transmitting data to the MUSIX software. All music notes will be transformed and sent to the analysis engine in MusicXML format.

3.2 Analyzing music data

Captured data are automatically sent to a database and then onwards to an analytics engine. The analytics engine offers a number of custom solutions specific to the requirements of each student or each class. Example analytics solutions include measure of students' level of understanding of music theory, models that mimic the challenges students faced in playing an instrument, and the capacity estimates on specific singing techniques.

MUSIX can detect, for example, a specific music theory topic that is not understood by a student, the tonic and the dominant notes that are well tuned for a student, or types of errors made by a student playing a piece in piano.

The analysis engine is the brain of the MUSIX system. It has the ability to parse the incoming data, at real time, and offer insightful information to the

teacher and students with respect to individual music skills. Analysis will be made on the level of confidence of a student pertaining to target skills.

3.3 Focusing on strengths and weaknesses

MUSIX has the ability to identify weaknesses of students as well as their strengths. Ethically valid monitoring of students gives them confidence because they know that the system will identify any misconception that needs timely remedy.

MUSIX, in addition to identifying weaknesses, also identifies positive advances students make at higher levels of granularity. Students tend not to see their performance at higher levels of abstraction. Their natural tendency is to tackle tasks that still need to be achieved and adjusted.

3.4 Giving feedback and guidance

MUSIX is valuable in precisely identifying student difficulties, conceptual issues faced by the student, and remedial activities that address the difficulties and issues. Students receive real-time feedback in relation to their answers to exercises, their play moves within a music game, or the quality of their notes in voice recordings. Continuous guidance is reassuring for students as they do not feel alone in their learning process but rather see that they are guided through specific steps and exercises to overcome their personal challenges.

If MUSIX detects the struggles of a student with respect to a specific learning facet, it has the means to offer alternative study pathways, in terms of additional hints, instructions, games, and quizzes. MUSIX lessons are designed based on the theory of experiential learning (Lee, Barker, & Kumar, 2011), allowing students to select the medium and mode of study that suit them, as depicted in Figure 1.

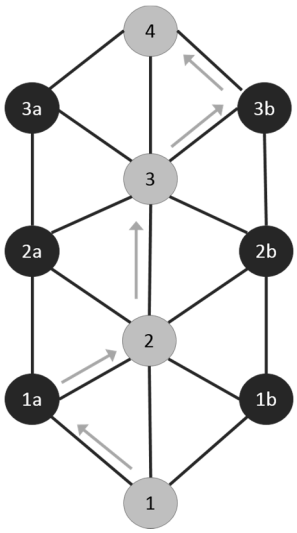


Figure 1 – Theory of experiential learning

3.5 Displaying interactive data

MUSIX has a dashboard that shows data being captured, explains the transformation of data, and interactively visualizes analysed data. Visual representation helps to locate study trends and gaps in the learning process. The dashboard is accessible to students, teachers, parents, and others involved in the music education development.

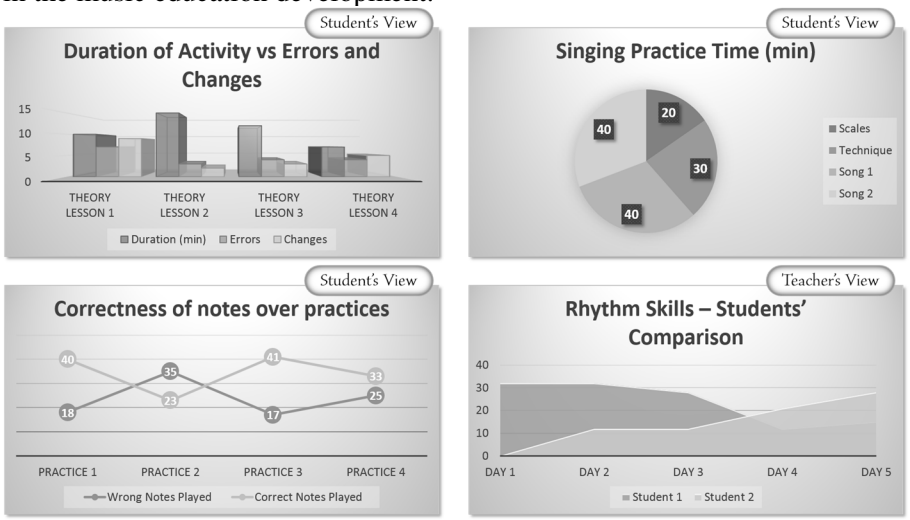


Figure 2 - Example of visualizations offered in the MUSIX dashboard

The teacher's view allows teachers to have an overall view of performances of their entire class at the same time, identifying students struggling with music concepts compared with the average performance of the class. The class average option is also available in the student's view to enable students compare their progress with classmates.

Overall, the student will be able to see, in real time, study material, concepts, and competences that have been mastered as well as gaps in their knowledge.

3.6 Encouraging self- and co-regulation

Self-regulation are co-regulation are features in MUSIX that target student motivation and positive engagement. In self-regulation, students set their own goals corresponding to a set of skills, identify study strategies to accomplish tasks leading towards mastery of these skills, measure the success of the goals, and adapt study strategies to accomplish the goals. These activities, setting of goals, picking of strategies, and executing and adapting plans, are termed as initiatives. Students can monitor their progress with respect to each initiative. Typically, students create initiatives as part of their interactions in the dashboard. The MUSIX system will continually and automatically update the parameters of each initiative, thus allowing students to engage explicitly with their own study.

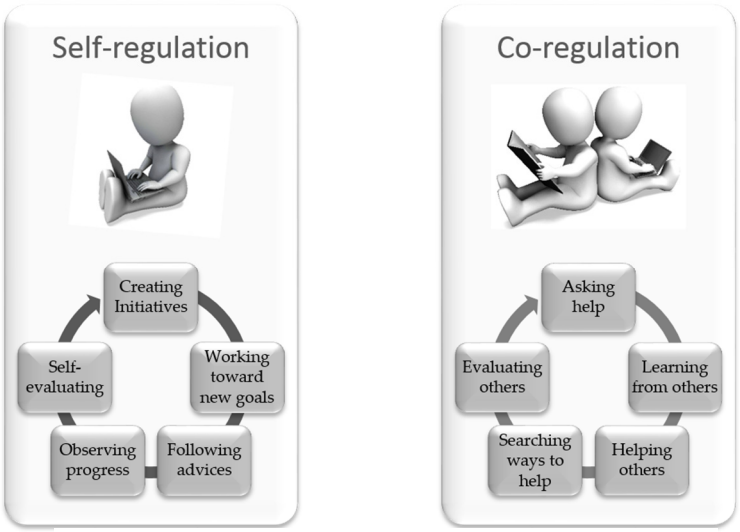


Figure 3 - Elements that are part of self- and co-regulation

Co-regulation is also a key feature in MUSIX where initiatives are created, monitored and modified by people other than the student. The social aspect of

co-regulation that involves friends, peers, teachers, parents, and even loved ones is important in music learning.

4. CONCLUSION

In order to get individual and collective insights on learner experiences in the music learning domain, this research applies learning analytics techniques in the music learning domain to define the learners' profiles by continually observing them during their learning process, to identify evidences of both misconceptions and competences, to determine their confidence in reaching the learning objectives, to provide richer insights to learners and instructors, and to track and promote self-regulatory traits of students. MUSIX, as a software agent, proves an effective and efficient complement to the human agent in his/her absence standing by the student to track information about the students' skills, challenges, weaknesses, successes, knowledge, attitude, and study habits and providing the teacher (human agent) with those valuable information even at a distance. The techniques applied in this chapter in MUSIX were also previously applied and experimented successfully in other learning domains such as programming, English writing, and mathematics.

MUSIX is providing teachers with in-depth information about the students' learning status that no traditional music teaching approach is able to provide on a constant basis. Music analytics allows teachers to be aware of the learning process adopted by students in each study session as well as in informal practice time. As discussed in this chapter, the group approach and punctual assessments do not allow to ensure the optimal learning experience for most students. MUSIX focuses on tracking relevant data about the students' daily routine in learning music in order to provide them with an individualized learning path based on the understanding that students' daily study habits is a key factor that affects their level of performance.

MUSIX teaching approach will impact the development of students as they go through instructional sessions, computer-based exercises, games, quizzes, social network interactions, and practice time in assessing what they have learned. In addition, MUSIX offers the ability to students to regulate their own learning habits. Whether it is music theory, vocal training or instrumental playing, MUSIX aims to contribute in the world of music education by tracking students' activities and explicitly capture the growth in music competence and music confidence thus improving the quality of future musicians.

REFERENCES

Aldalalah, O. A. (2010). Music Intelligence and Music Theory Learning: A Cognitive Load Theory Viewpoint. *International Journal of Psychological Studies, 2*(2), 150-158.

Bauer, W. I. (2014). Music Learning and Technology. *New Directions In Music Education*.

Beatty, R. J. (2015). Contemporary Practices of Assessment and Evaluation in Music Education. In S. Schonmann, *International Yearbook for Research in Arts Education* (pp. 419-424). Germany: Waxmann Verlag GmbH.

Guillot, C. (2015). *MUSIX*. Retrieved from Learning Analytics: http://learninganalytics.ca/research/?page_id=1845

Lee, S., Barker, T., & Kumar, V. (2011). Learning Preferences and Self-Regulation – Design of a Learner-Directed E-Learning Model. In T.-h. Kim, H. Adeli, H.-k. Kim, H.-j. Kang, K. J. Kim, A. Kiumi, & B.-H. Kang, *Software Engineering, Business Continuity, and Education* (pp. 579-589). Berlin: Springer Berlin Heidelberg.

McPherson, G. E., & Renwick, J. M. (2000). Self-Regulation and Musical Practice: A Longitudinal Study. *ICMPC*.

Rodrigues, A. N., & Dos Santos, S. C. (2013). A Systems Approach to Managing Learning based on Bloom's Revised Taxonomy to Support Student Assessment in PBL. *2013 IEEE Frontiers in Education Conference (FIE)*, (pp. 1004-1010).

Wise, S., Greenwood, J., & Davis, N. (2011). Teachers' use of digital technology in secondary music education: illustrations of changing classrooms. *British Journal of Music Education, 28*(2), 117-134

Chapter 8

SEMI-SUPERVISED PRODUCT SPECIFICATIONS EXTRACTION FROM THE WEB

George Krys[1], Ebrahim Bagheri[2]
[1]*School of Computing and Information Systems, Athabasca University*
[2]*George Vari Engineering and Computing Centre, Ryerson University*

Abstract: This paper introduces an approach that utilizes the inductive semi-supervised learning strategy – self-training model for extracting property-value pairs from a collection of Web pages. The proposed work employs a novel concept of short properties and values for learning high confidence property-value pair seeds. The seeds are then used to discover repetitive HTML formatting patterns and consequently, using these patterns as the wrappers to extract the rest of the property-value pairs that the Web page contains. The experimentations on the collection of Web pages, drawn from over one hundred diverse, real life electronic goods retailer Web sites, show promising results.

Key words: Learning — Knowledge Acquisition, Semi-Supervised Machine Learning, Web Information Extraction, Product Property Name and Value Pair

1. INTRODUCTION

With billions of dollars already spent by digital shoppers each year and favorable forecasts of capturing even larger share of sales in the near future, online retail is one of the fastest growing industries. The success of e-commerce in turn has given rise to online aggregators, search engines, comparison shopping sites, shopping portals, deal and auction sites that help users sift through the retail debris. However, each online aggregator faces the

extremely challenging problem of building a cohesive product catalog, which can be leveraged for smarter search and better customer experience.

The information about products is very volatile, i.e., they are frequently updated and typically generated dynamically on the fly when a user wants to see the product specifications. The specification Web pages vary in HTML formatting and are often only optimized for human browsing rather than machine processing. There are several irregularities in the presentation of product property-value pairs (PVPs) on numerous retailers' Web sites, such as layout formats that range from regular table, list to unstructured free text, usage of different product properties or different names of the same properties or values (i.e., synonymous words or phrases having the same or similar meaning, abbreviations), different ordering of properties from site to site as well as omission of some PVPs or values even on the same site. Moreover, the product detail pages, besides product specifications contain other blocks of information such as toolbars and navigation bars that contain links for browsing the Web site inventory, additional information about the product, customer reviews, among others. Although these supplementary informational blocks can enhance consumer appeal, usability, and visual attractiveness, it brings great challenge for locating the PVPs on the page in order to automatically extract them.

In light of these challenges, the key objective of our work is to develop an efficient algorithm capable of tackling the presentation irregularities and specification block identification obstacles of product specification pages for extracting PVPs. We employ a combination of semi-supervised machine learning and pattern mining techniques for discovering and extracting property-value pairs from product specification Web pages. In our self-training machine learning model, we introduce the novel concept of short properties and values for discovering high confidence PVP seeds. The seeds are then used to discover repetitive HTML formatting patterns found surrounding the PVPs. These discovered patterns serve as the wrappers to extract the rest of the property-value pairs that the Web page contains.

2. RELATED WORK

The procedure for extracting content from a Web page based on the knowledge of its format is often referred to as 'wrappers' and the process of information extraction (IE) is known as Wrapper Induction (WI). Formally, a wrapper is a function for extracting the related content from a Web page while discarding the irrelevant text (Kushmerick, 2000). The main wrappers' assumption is that one can find repetitive patterns in one or more Web pages that conform to a common template. The goal of WI is to automatically

generate a wrapper that is used to extract the targets from an information resource.

In a supervised wrapper induction (Kushmerick, 1997; Wang & Hu, 2002), a set of extraction rules are learnt from a set of manually labeled pages that are used to extract data items from similar pages. Because manually labeling pages is labor intensive, much effort has been devoted to automating the wrapper generation process by employing semi-supervised or unsupervised machine learning (ML) techniques (Chang et al., 2006).

IEPAD (Chang & Lui, 2001) is one of the first IE systems that semi-automatically generalizes extraction patterns from unlabeled Web pages. The system identifies record boundaries by repeated pattern mining and multiple sequence alignment utilizing a PATRICIA trie data structure (Gusfield, 1997; Morrison, 1968). However, the method described in (Chang & Lui, 2001) produces a large number of irrelevant patterns, and in several cases fails to find the right one. It mainly occurs because the domain-specific information is not utilized in this approach. To deal with this issue, IPEAD has an interface for users called pattern viewer through which they can choose a proper record extraction pattern. In contrast, our method integrates domain-specific information into the learning mechanism, which aids in the discovery of the relevant patterns and hence requires much less direction from the user.

Some of the most cited unsupervised systems include RoadRunner (Crescenzi, Mecca, & Merialdo, 2001) and ExAlg (Arasu & Garcia-Molina, 2003). These systems compare HTML pages and generate a wrapper based on their similarities and differences. The process, starts by taking one input page as an initial version of the wrapper, then it is matched against the test page and iteratively refined solving mismatches by generalizing the wrapper. More recently, comparable unsupervised adaptive template-based method was proposed in (Tang et al., 2012). The method includes constructing the attribute word bag using Web titles from a single domain. The bag of words is then used to learn high-quality page templates by selecting most frequent patterns across multiple pages. However, all these methods require sufficient number of equivalent pages to discover their common templates. Our algorithm, in comparison, is template-independent; that is, given a small dictionary of short properties and values, the product's PVPs can be extracted even from a single document.

Similar template-independent but unsupervised approach is proposed in (Walter, 2012). The approach combines several different features from a document's tokens, its tree representation, and visual information for clustering Web page elements. The element lists are then purged to drop insignificant clusters and create extraction candidates. Finally, based on the best-rated candidate a wrapper consisting of a set of XPath queries is created and executed to extract actual product specifications. However, the main

drawback of this approach is the quite high costs in terms of time. Even though during the algorithm evaluation the extraction routines that exceeded 100 seconds were terminated, an average runtime to process a Web page was 13 seconds, which makes the method infeasible in practice.

The pivotal point of our research was the algorithm for constructing document-specific character-level wrappers automatically proposed in (Wang & Cohen, 2009). The algorithm finds maximally long contextual strings that bracket at least one seed instance of every seed using PATRICIA tries. Although the main focus in (Wang & Cohen, 2009) is on the expansion of a partial set of "seed" objects into a more complete set by extracting named entities with wrappers, we adapted this method for binary relations (i.e., property and value pairs) extraction and received very positive results.

Lastly, our work bears a resemblance to the system proposed in (Wu et al., 2009). The paper describes a semi-supervised, template-independent method, which uses a few manually labeled pages for a product domain. The labeled pages are combined with unlabeled ones to boost the learning of candidate attributes using a co-training algorithm with Naïve Bayes classifier. The candidates are used to identify product specification blocks on the page and to extract data of interest. However, unlike our approach, the newly identified name-value pairs (NVPs) are only temporarily added to the training set. They are removed from the training set when the classifier moves to the next page. It generally means that for each page the learning starts from scratch (i.e., original training set) and the process is not enhanced with new knowledge after it progresses to the subsequent pages. As stated in (Wu et al., 2009), the reason for this is that the misclassified text nodes cause semantic drift and lead to many invalid extractions. Our approach solves the semantic drift problem by utilizing only the high confidence short property and values. It also progressively increases the number of learned PVPs with each processed page, which, in turn, leads to higher precision and recall.

3. PROPOSED TECHNIQUE

In order to discover patterns and extract PVPs from product specification Web pages, we propose a semi-supervised, template-independent method. As illustrated in Fig. 1, the method comprises four main tasks: preprocessing, seed learning, pattern discovery, and pattern-based extraction.

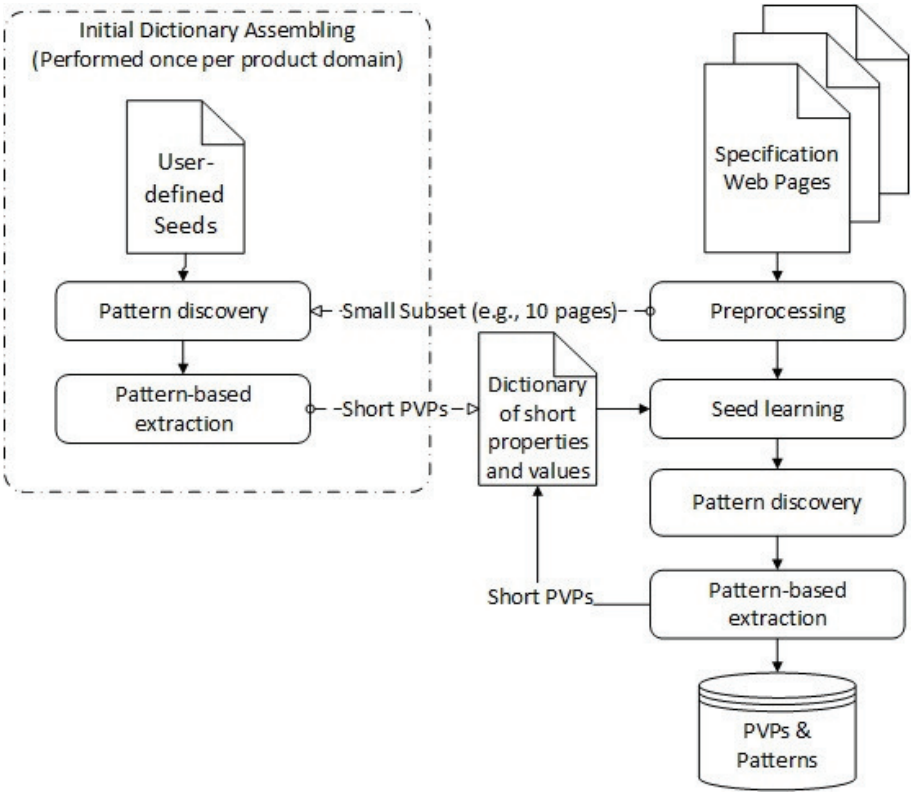


Fig. 1. The proposed semi-supervised property-value pair extraction algorithm.

The following subsections present the details about the four steps of the proposed technique.

3.1 Preprocessing

To reduce the amount of characters that need to be processed by the character-level pattern discovery algorithm, the input HTML document is parsed chunk-by-chunk and converted to a clean source document. The chunks are the open tags, closed tags, and the text between the tags, which are referred to as the text chunks in this paper. A text chunk may contain irrelevant text (i.e., neither property nor value, nor PVP) or either property or value, or PVP. According to the relative position of a property and value, we assume that a Web page may meet two conditions: i) the product property name and value, which constitute a PVP, organized in different but adjacent text chunks (e.g., <tr><td>Camera Resolution</td><td>3.5 megapixels</td></tr>); ii) the product property name and value belong to the same text chunk and separated by a symbol such as the colon symbol ":" (e.g., Camera Resolution: 3.5 megapixels).

The preprocessing task includes removing new line symbols, tabs, and extra white spaces as well as tags that certainly do not surround PVPs (e.g., header, script, image, etc.) and contents between these opening and closing tags. To ensure presence of repetitive patterns in HTML formatting, we also remove unique attribute values (i.e., id and name) and strip any other attributes and their values of all the symbols, white spaces, and numbers (e.g., <th scopecolclassspecvaluecol>). The text chunks are checked for a separator (e.g. ':') and only the first separator in the chunk is retained if there is more than one separator in the text chunk. Additionally, we automatically close <tb>, <th>, and <li> tags if they do not have corresponding closing tags.

3.2 Seed Learning

One of the main steps of our technique is to identify a few (at least two) PVPs that can be used to learn the wrapper for the rest of the PVPs. We refer to these base PVPs as seeds. To automatically learn the seeds and discover patterns, we use the following two basic assumptions: (1) a property always comes before a value, and (2) the Web page uses some repetitive formatting for presenting PVPs. If there is a repetitive pattern, then only a small number of PVPs are required for its discovery.

Our approach to seed learning employs the inductive semi-supervised machine learning strategy – self-training model and is based on the observation that there is a limited number of common short property-value pairs in any product domain. (e.g., Weight: 1 kg, Display Resolution: 1024 × 768). Because of the limited number of common short PVPs and very little variety in their unique spelling, no typical natural language processing (NLP) tasks were required (i.e., POS tagging, stemming, etc.).

3.2.1 Dictionary of Short Properties and Values

The automated seed learning process is preceded by the initial semi-automated assembly of the domain dictionary of short properties and values. The semi-automated process is shown in the dashed box in Fig. 1. The process uses a small subset of pre-processed product specification pages from a specific domain (e.g., 10 pages) and user-provided seeds to automatically discover patterns and extract PVPs. The extracted short PVPs during this stage are saved to the dictionary and serve as the initial feature set for the semi-supervised machine learning algorithm of the main process.

We tested two types of domain dictionaries (aka feature types): i) Bag-of-Words (BOW), and ii) n-grams (NGRAM). The BOW dictionary contains words found in short properties and values disregarding the word order. The NGRAM dictionary (i.e., bigram, trigram, …, n-gram where n is the number

of tokens) contains entries of a property or value text string without the white spaces between the words (e.g., CAMERARESOLUTION); thus preserving the word order. Both of the dictionaries maintain the number of each unique word or n-gram occurrence in the training set.

3.2.2 Classifiers

To automatically discover the seeds in the page, we experimented with two advanced classifiers, namely Bayes Point Machine (BPM) and Support Vector Machine (SVM) as well as with a simple probabilistic Naïve Bayes classifier. The initial domain dictionary of short properties and values entries (tokens) serve as the set of features and the documents that the dictionary is built on – as the training data for the classifiers. The classifiers are to predict P (Property) or V (Value) labels for each text chunk of the tested Web page whose number of words do not excessed the token limit. The token limit is the maximum number of tokens (i.e., number of words) for property and value pairs to be accepted as a short PVP.

For BPM and SVM classifiers, a vector whose dimensions are equal to the number of features (i.e., number of property and value entries in the dictionary) represents the test data. The BPM classifier in our experiments uses the Bernoulli model of features (i.e., binary occurrence information) ignoring the number of occurrences. The train\test text string is represented as one row of comma separated binary values where the value 1 marks the presence of a word (i.e., feature) in the train\test text string. The position within the string indicates an ordinal position of the entry (i.e., feature) in the dictionary. The SVM classifier, on the other hand, takes into consideration the number of occurrences of a feature in a training set. The tested text string for SVM is represented similarly to the following string of values: 1:3,32:1,56:45, where the first number is the ordinal value of the feature and second is the number of occurrences of the word in the training set. Since both, BPM and SVM assign one of the labels to all tested text chunks, we experimentally determined the classifiers' cut off scores to separate the high and low confidence predictions.

The Naïve Bayes (NB) classifier matches individual tokens of the provided text chunks (i.e., test data) with the dictionary of short properties and values. It calculates the probability of the text chunk being a property or value, then performs the sorting operation on the probabilities. The first element in the sorted list of probabilities (i.e., highest probability) reveals the label to assign. However, if the probabilities of the text chuck belonging to one of two categories are the same, the classifier assigns U (Undetermined) label; in other words, the classifier cannot decide which category it belongs to because of a

lack of information. This "naivety" allowed accepting any text chunk labeled as P or V as the high confidence predictions of the NB classifier.

Lastly, after the tested text chunks were labeled by any of the above classifiers, a text chunk classified as property was selected as a part of a seed only if it is directly followed by the text chunk classified as value.

3.3 Pattern Discovery

A pattern (aka wrapper) in our proposed approach is a subclass of regular expressions over an alphabet of tokens. An example of a pattern is shown below:

<tr><td>(.*?)</td><td>(.*?)</td></tr>

where <tr><td>, </td><td>, and </td></tr> are tokens that match groups of HMTL tags and the (.*?) is a wildcard that matches any string of characters (i.e., including HTML tags). The following string matches the above pattern:

<tr><td>Camera Resolution</td><td>3.5 megapixels</td></tr>

To discover patterns, we first find all instances of the identified seeds (from Section 3.2) within the document of interest. The instances are grouped by the context that separates properties and values. The middle context as well as three-character long left and three-character long right context is maintained with each instance. Table 1 shows an example of seed instances found for one of the corpus documents.

Table 1. Example of seed instances.

Left	Property	Mid	Value	Right
ol>	Sensor Type	</td><td classcol>	CMOS	</t
ol>	Focal Length	</td><td classcol>	8.50 mm	</t
ol>	Color	</td><td classcol>	Black	</t
li>	Sensor Type	:	CMOS	</l

In the next step, we filter out the seed instances that do not have at least one counterpart with the same left, middle, and right context (i.e., Sensor Type: CMOS in our example), then join each PVP instance together using the middle context (e.g., Col-or</td><td classcol>Black). Next, for each seed instance, the system extracts a left character string, which starts from the first character of the document and ends before the first character of the joint PVP and a right character string, which starts from the first character after the joint PVP and ends at the last character of the document. The left character strings is inserted into the left-context PATRICIA trie. The right character string is inserted in reverse character order into the right-context PATRICIA trie. Every node in both tries maintains a list of IDs for keeping track of the seed instances that follow the string associated with that node.

As a final step, the tries are used to find maximally long contextual strings that bracket at least one seed instance of every seed. Similar to (Wang & Cohen, 2009), we first find all the longest possible strings from one trie, then for every such string, find the longest possible string from another trie such that both of them bracket at least one occurrence of every given seed in the document. Patterns are assembled for PVP extractions using the found left and right contextual strings and the middle context of the group of seeds

3.4 Pattern-based Extraction

It should be noted that the longest character-level prefixes and suffixes found in the method proposed in Section 3.3 are not always "HTML-compliant" and some-times produce patterns similar to the following example:

/td></tr><tr><td width>(.*?)</td><td>(.*?)</td></tr><tr><td width>

However, one of the major assumptions of our approach is that there has to be at least one HTML pattern (i.e., HTML-compliant pattern) in order to extract PVPs. Therefore, our algorithm attempts to normalize patterns by removing all the incomplete tags (i.e., /td>), tags that have no matching opening or closing counterparts, and tags that surround other tags, such as table and table row tags, which surround table cell tags (i.e., < table>, <thead>, <tbody>, <tr>), and list tags which surround list items tags (i.e., <ul>, <ol>, <dl>). The invalid HTML pattern example above would be normalized to the following pattern: <td width>(.*?)</td><td>(.*?)</td>

The normalized (if required) pattern is used for PVP extraction. The regular expression is applied to the document of interest in both directions: left-to-right and right-to-left, then the two groups of regular expression matches are used to assemble property-value pairs. The reason for two-directional application is that in some cases, specifically with patterns based on , <p>, or HTML blocks, the first property or the last value of the extracted PVPs may include extra markup and text. When left-to-right and right-to-left extracted PVPs do not match, our system validates the inner HTML of non-matching items and selects shorter strings that do not contain any markup or longer strings only if they have valid HTML.

After all PVPs are extracted, the relative short PVPs (e.g., three-word property and three-word value) are added to the dictionary and their counts of occurrences are increased. These PVPs are used in the feature set for the subsequent classification. Finally, the discovered patterns and extracted PVPs are saved to the database for product specification cataloging.

4. EVALUATION

The evaluation of the proposed approach was performed using the standard IE performance measures of precision, recall, and F_1 score. The corpus used in the experiments was collected from over one hundred distinct electronic goods retailers' Web sites from two domains – Digital Cameras and Smartphones. In our experiments, we focused on the ability of the algorithm to discover broad variety of patterns and therefore selected specification Web pages that are uniquely formatted.

The entire corpus was semi-automatically processed; and then manually examined, edited and together with missing PVPs combined into *page item set* – a set of actual PVPs that the document contains. The page items of each document were saved to the database and served as the ground truth test data set in order to benchmark our proposed algorithm. Ten documents from each product domain and their page items also served as the seeds for performing the semi-supervised learning.

The experiments were performed for two types of PVP extractions: (1) utilizing known seeds (i.e., user provided seeds), in order to assess effectiveness of the pattern discovery functionality, and (2) utilizing learned seeds, in order to identify best suitable ML classifier and corpus processing strategy.

4.1 Pattern Discovery Evaluation

To evaluate the effectiveness of the developed PVPs pattern discovery technique, the accuracy of the technique was measured on the gold standard corpus. To infer the unbiased performance of the algorithm, its precision, recall, and F_1 scores were calculated for exact matches (i.e., actual PVP matched expected PVP word-for-word) and for partial matches (i.e., actual PVP was close enough to the expected PVP but contained some additional, irrelevant words or had some words missing).

The results of our tests showed that the character-level approach for constructing document-specific wrappers works very well for discovering patterns and extracting property-value pairs from Web pages. Its application allows discovering patterns not only for properties and values structured in simple tables and lists but also uncommon, unique PVP formatting (i.e., patterns based on or HTML blocks) similar to the ones shown in Table 2.

Table 2**.** Examples of discovered patterns

URL	http://estore.canon.ca/webapp/wcs/stores/servlet/prouct_12152_10102_10355_24
Pattern	(.*?) (.*?)

URL	http://www.naaptol.com/digital-cameras/olympus-point-shoot-vg-150/p/3683646.html
Pattern	<div classtechdsclayout><div classclm> (.*?)<span classhelpicon><div idclasshelpdatastatehiddenstyledisplaynonepositionabsolute> <div classclm> (.*?)
URL	http://store.sony.com/xperia-z1-zid27-xz1c6902//cat-27-catid-all-unlocked-phones?vva_colorcode=000000&_t=pfm%3dcategory
Pattern	<h5 classwsproducttitlefndatadynamicblockiddatadynamicblockname>(.*?)</h5><div classwsgroupsnyproductrating><div classwsgroupcontentssnyproductratingcontentsstylewidthpx><div\ classwsgroupsnyratingstars><div classwsgroupcontentssnyratingstarscontents> <div classsnyavgrating>(.*?)

Table 3 presents the final results of the experiments. In the first row of the table the partial match extractions were counted as both – False Positive and False Negative extractions; in the second, they were counted as True Positive extractions. As the test results indicate, the implemented algorithm is very effective in discovering patterns for PVP extraction.

Table 3. Evaluation results of the pattern discovery algorithm performance utilizing known seeds.

	Expected	**Actual**	**TP**	**FP**	**FN**	***Precision***	***Recall***	F_1
Exact	4848	4841	4790	51	60	99.34%	98.90%	99.11%
Partial	4848	4841	4827	15	22	99.77%	99.37%	99.55%
TP – true positive, FP – false positive, FN – false negative.								

4.2 Seed Extraction Evaluation

The objectives of the second set of experiments was to find the most suitable ML classifier for identifying PVPs and determine the corpus processing strategy, which will deliver the highest precision, recall, and F_1 score. To meet the former objective, we integrated and extensively tested the following classifiers: i) Microsoft Infer.NET Bayes Point Machine [1], ii) .NET implementation of LIBSVM[2], and iii) a custom implementation of a simple Naïve Bayes classifier.

To find the best-suited classifier for the task, we compared the performance of BOW, NGRAM and BOTH feature types for each classifier. The BOTH type works in the following fashion: first the system attempts to find seeds using the NGRAM approach, then, if it does not succeed (i.e.,

[1] http://research.microsoft.com/en-us/um/cambridge/projects/infernet/
[2] http://www.matthewajohnson.org/software/svm.html

number of found seeds is less than 2 and only then, it uses BOW approach. The chart in Fig. 2, shows that the BOW results in higher recall and NGRAM in higher precession as well as that all classifiers deliver identical results with BOTH feature types setting. The BOTH approach smoothens the difference between BOW and NGRAM precision and recall and delivers highest F_1 score. Therefore, since all classifiers with BOTH settings deliver identical precision and recall, the only performance indicator that discriminates them is the task execution time: around 8 min for BPM, 7 min for SVM, and 2 min for NB on commodity hardware with the in-house system optimized for research (i.e., logging and output to the screen of each processing step during the test) rather than efficient performance. Thus, NB classifier was selected as the best suited classifier for the task because of the shortest processing time.

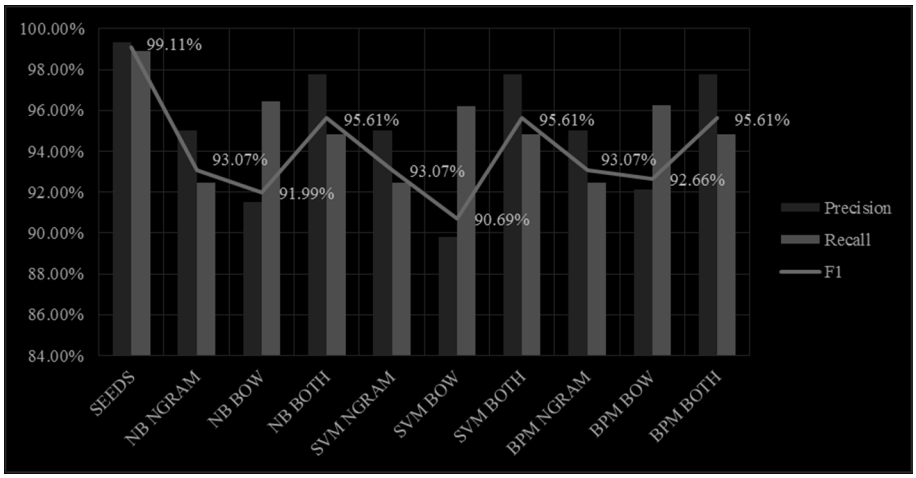


Fig. 2. Feature type evaluation results.

After the classifier was selected, we conducted an experiment to determine the corpus processing strategy. In this experiment, four different strategies were compared with each other as well as with the corpus processing results utilizing known seeds. The four strategies as follows:

1. *One iteration NB BOTH plus an extra iteration with BOW features and increased token limit.* The NB BOTH processing was enhanced with an extra iteration in which the failed documents (i.e., documents that did not produce any extractions) were processed with BOW features and the token limit was increased by two. The purpose of increasing the token limit is to successfully process the Web documents that might not have enough 3 token PVPs (default) for discovering patterns (i.e. less than two short PVPs per page).
2. *One iteration NB BOTH with 20 training set pages plus an extra iteration with BOW features and increased token limit.* The same arrangement as in

the previous strategy but the training set was composed of 20 corpus documents that contained the maximum number of PVPs. The aim was to test if changing the training data will have effect on the processing results.

3. *Two iteration NB BOTH plus an extra iteration with BOW features and increased token limit.* The first iteration performed with the default settings. In the second iteration, all documents are processed using features collected during the first iteration and only features that do not already exist in the dictionary of short properties and values were added to the list.
4. *Three NB NGRAM iterations plus an extra iteration with BOW features and increased token limit.* The aim of this strategy was to test if using primarily NGRAM features will reduce the number of erroneous extractions. The second and extra iterations were similar to the ones used in the previous strategies. In the third iteration only the failed pages were processed using BOW features and no features were added in this iteration to the dictionary of short properties and values.

The chart in Fig. 3 illustrates that the strategy number three (two iterations) delivers the best results. We believe that the proposed technique can easily be extended to include other product domains. However, greater variety of domains needs to be used in order to get more conclusive results.

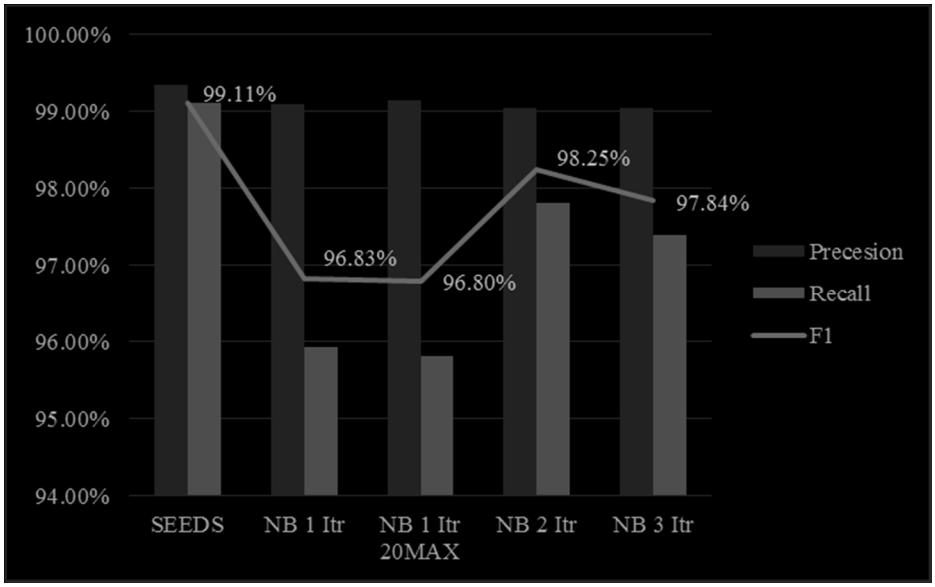


Fig. 3. Corpus processing strategies evaluation.

5. CONCLUSIONS

In this paper, we have proposed an algorithm for extracting property-value pairs from a collection of Web pages. The algorithm employs inductive semi-supervised learning strategy – self-training model and a novel concept of short properties and values for learning high confidence property-value pair seeds. The seeds were then used to discover repetitive HTML formatting patterns and consequently, using these patterns as the wrappers, extract the rest of the property-value pairs that the Web page contains. The key feature of this approach is the focus on semi-supervised learning of a limited number of short property-value pairs per product domain, which normally do not vary in spelling. It greatly reduces the annotation effort and amount of data (i.e., Web page content) that needs to be processed, streamlines the ML task because no typical NLP preprocessing is required (i.e., POS tagging, stemming, etc.), and allows usage of simplistic classifiers such as Naïve Bayes, all of which in turn makes the entire process very efficient. The empirical testing on the collection of Web pages, drawn from over one hundred diverse, real-life electronic goods retailers' Web sites indicate that the algorithm performs well.

REFERENCES

Arasu, A., & Garcia-Molina, H. (2003, June). Extracting structured data from Web pages. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of data* (pp. 337-348). New York, NY: Association for Computing Machinery.

Chang, C.-H., Kayed, M., Girgis, M., & Shaalan, K. (2006). A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering, 18*(10), 1411–1428. doi:10.1109/TKDE.2006.152

Chang, C.-H., & Lui, S.-C. (2001, April). IEPAD: information extraction based on pattern discovery. In V. Y. Shen (Ed.), In *Proceedings of the 10th International Conference on World Wide Web* (pp. 681–688). New York, NY: Association for Computing Machinery.

Crescenzi, V., Mecca, G., & Merialdo, P. (2001, September). RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *Proceedings of the Twenty-seventh International Conference on Very Large Data Bases* (pp. 109–118). San Francisco, CA: Morgan Kaufman Publishers Inc.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. New York, NY: Cambridge University Press.

Kushmerick, N. (1997). *Wrapper induction for information extraction* (Doctoral dissertation). University of Washington

Morrison, D. R. (1968). PATRICIA - Practical Algorithm To Retrieve Information Coded in Alphanumeric. *Journal of the ACM*, *15*(4), 514–534. doi:10.1145/321479.321481

Tang, W., Hong, Y., Feng, Y.-H., Yao, J.-M., & Zhu, Q.-M. (2012). Simultaneous Product Attribute Name and Value Extraction with Adaptively Learnt Templates. In *Proceedings of 2012 International Conference on Computer Science and Service System (CSSS)*, (pp. 2021–2025). Washington, DC: IEEE.

Wang, R. C., & Cohen, W. W. (2009, August). Character-level analysis of semi-structured documents for set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*: *Volume 3* (pp. 1503-1512). Morristown, N.J: Association for Computational Linguistics.

Wang, Y., & Hu, J. (2002, May). A machine learning based approach for table detection on the Web. In *Proceedings of the 11th international conference on World Wide Web* (pp. 242–250). New York, NY: Association for Computing Machinery.

Walter, M. (2012, November). Unsupervised Extraction of Product Information from Semi-structured Sources. In *Proceedings of the 13th International Symposium on Computational Intelligence and Informatics* (pp. 257–262). Washington, DC: IEEE Computer Society.

Wu, B., Cheng, X., Wang, Y., Guo, Y., & Song, L. (2009, September). Simultaneous Product Attribute Name and Value Extraction from Web Pages. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (pp. 295–298). Washington, DC: IEEE Computer Society

Chapter 9

USER MODELING FOR COURSE PLANNING AND SCHEDULING

Fuhua Lin, M. Ali Akber Dewan, Alex Newcomb
School of Computing and Information Systems, Athabasca University

Abstract: Course offering determination (COD) is a complex task for the educational institutions by which they decides what subset of courses an academic department or program should offer in a given academic term or semester. For an effective COD, historic data in enrollment, constrained in budget, staff, and resources, and student's course selection preferences and priorities are needed to be taken into account. A poorly designed COD may lead to a low enrollment to the program, delayed graduation, and students increased dissatisfactions. A multi-agent based framework to facilitate COD is proposed by the authors in (Lin & Chen, 2013), which uses Contract Net Protocol, Single Transferable Voting, and Monotonic Concession Protocol. The multi-agent system (MAS) consists of an administrator (AD) agent, a group of student (SA) agents, and a student representative (SR) agent. In this paper, the modeling and implementation details of the SA agents are presented. The prototype of the system is implemented using Java Agent Development Framework (JADE). The system is expected to solve the problems in optimization and with flexibility in a fair and rational way, which can balance the competing needs of academic requirements, economics, and student preferences.

Key words: Multi-agent systems, agent-oriented software engineering, JADE, course-offering determination, Contract Net Protocol.

1. INTRODUCTION

Course offering determination (COD) is a process of deciding what courses of an academic program or school will be offering for the upcoming one or

more semesters (Lin, Newcomp, & Armstrong, 2012). There are many factors that an administration needs to take into account in order to provide an effective list of offering courses. These factors include historic data in enrollments, budget, staffing and resource constraints, students preferences and priorities. A department of an academic institution may not be able to offer all courses in a program every semester, especially under contemporary fiscal and staffing constraints. However, a poorly designed COD can lead to several problematic situations, such as low enrollment, delayed graduation of the students, and students increased dissatisfactions. Some courses could only be arranged every other semester or even less frequently.

In the current course offering workflow, when the registration period for a new semester approaches, a course delivery schedule becomes available and the student can select courses to be taken in the coming semester. The competing or even adversarial goals of students and the department as well as the mutability of those goals indicate that COD is a complex constraint-satisfaction problem. Effective COD permits the efficient assignment of limited resources like faculty, labs, and classrooms, while satisfying the desires of most students.

Multi-agent system (MAS) allows the representation of every principal in a system as a single autonomous agent with unique goals and permits decision-making based on the preferences of multiple agents (Weiss, 1999) (Conitze, 2010). The MAS approach can be used to solve the constrained-satisfaction problem of COD because of the following reasons — (i) optimal solution of this problem can be changed during run time; (ii) relation between a user and the scheduling system lasts for a long period of time, which increases the possibility of learning by feedback; (iii) COD is a time consuming and tedious task using manual process; (iv) multiple-parties are involved in CODs (i.e., program administrators and students), all of which are required to be satisfied at least a minimum level from the provided solution; and finally, (v) unpredictable job market and student preferences needs to be taken into account fast and flexibly to environmental variables and their changes. Since the goals of the students and program administrators are different, therefore, there are conflicts of interests between them. These conflicts should be resolved in a fair cooperative decision making manner. The main research question of COD that is to be solved using MAS is "*what COD strategy of a program in an institution maximizes the satisfaction of the students and the enrollment of the courses within institutional constraints (i.e., limited budget, staff and teaching resources)?*"

In (Lin & Chen, 2013), authors modeled COD as a multi-agent constraint resource allocation problem and designed a mechanism to identify optimal solutions using voting and agent negotiation. Various theories of Agent-Oriented Analysis and Design (Wooldridge , Jennings, & Kinny, 2000) and

Agent-Oriented Software Engineering (AOSE) (Winikoff & Padgham, 2013) have been applied in modeling this framework. In agent-based recommendation applications, users need to feel easier and more comfortable to express their goals for obtaining items rather than specifying the features of the items. In view of this, this multi-agent framework is consisted with three different types of agents – Student Agent (SA), Student Representative Agent (SR), and Administrative Agent (AD), where the relations among these agents are modeled using contract net protocol (Smith, 1980), single transferable voting (Brams & Fishburn, 2002), and monotonic concession protocol (Endriss, 2006). Since, the overall system is very large and complex, this paper mainly focus on modeling and implementation details of SA within JADE and JASON framework.

The organization of this paper is as follows. A brief literature review relevant to this work is presented in Section 2. Section 3 provides details of SA modeling and implementation. Experimental results are presented in Section 4. Finally, the paper ends with the concluding remarks.

2. LITERATURE REVIEW

Recommender systems have gained considerable interest since the 1990s as a means of helping users to deal with ever-increasing problems of information overload (Resnick & Varian, 1997) (Burke, 2002). In the field of education, researchers have used different recommendation techniques for different purposes, such as suggesting online learning activities or optimum browsing pathways to students (Farzan & Brusilovsky, 2006) (Tang & McCalla, 2005), making recommendations to courseware authors (Garcia, Romera, & Castro, 2009), and providing advice to high school students and college freshmen that are seeking a potential major (Grupe, 2002). There are three primary approaches for computing recommendations: content-based techniques, collaborative filtering, and demographic techniques. Content-based techniques rely on the availability of descriptive metadata that captures the essence of the items available for recommendation and compute similarities between items by comparing item characteristics. Collaborative filtering techniques provide an alternative strategy that replies on rating-based user profiles instead of descriptive meta-data (Schafer, Frankowski, Herlocker, & Sen, 2007) (Smyth & Cotter, 2001). Demographic techniques make recommendations based on demographic classes, by analyzing personal attributes of the user (Krulwich, 1997). Likely items for recommendation are identified because clusters of users with similar personal attributes have demonstrated similar needs or tastes.

COD is a kind of recommender system, where MAS has been proven to be as an effective solution (Lin & Chen, 2013). However, the majority of such works have been in the area of providing learner support and developing learning environments (Smyth, Shang, Shi, Chen, & Shing, 2001). In fact, the best sources of good research on applying MAS to problems of planning and scheduling, preference elicitation, and negotiation may be found outside the educational realm in sectors such as manufacturing (Shen, 2002) and health care (Kirn, Herzog, Lockemann, & Spaniol, 2006). The potential effectiveness of decision support system and other AI systems in supporting administration in an increasingly complex education sector have also been recognized (Kannan & Kasmuri, 2005). There has been a relative paucity of works focusing in the role of advisor or administrator in providing educational resources and support save for work on problems with a tradition of AI applications, such as automated course scheduling and computer-aided academic advising (Opera, 2007).

Hamdi (Hamdi, 2006) proposed a system Masacad that tackles the program planning problem using MAS approach and used neural network aiming at finding the correct agent function. It is a multi-agent information customization system that adopts machine-learning paradigm to advice students by mining web information. Vainio and Salmenjoki (Vainio & Salmenjoki, 2005) proposed an agent-based approach to designing and updating a personalized study plan in collaborative environment. The system is based on learning agents able to suggest a study plan and if needed identify potentially problematic choices in the future, thus bring dynamics in the system. It shows that collaborating with other student agents in a multi-agent environment, the chances of finding a mutually beneficial result is improved. Bruns (Bruns, 2006) presents software architecture of a new generation of advisory systems using intelligent agent and semantic web technologies. To the best of our knowledge, the system proposed in (Lin & Chen, 2013) by the authors of this paper is the only system that addresses COD problem within a MAS framework.

3. SYSTEM ARCHITECTURE

The overall architecture of the proposed system for COD based on MAS is shown in Figure 1. The proposed MAS system consists of an administrator agent (AD), a student representative agent (SR), and a group of student agents (SAs). This system correlates with an academic program, where a course scheduling process is initiated by having the program administrator determined the priority of courses available in the program based on

expressed student needs, preferences, and goals. The agents have distinct areas of concern and intent, but collectively interact to generate a set of recommendations for courses to be offered that will be satisfactory to most students while fitting within the operational limitations of the offering program. Considering the large scale and complexity of the whole system, we have limited our scope on modeling and implementing the SAs of the proposed system in this paper.

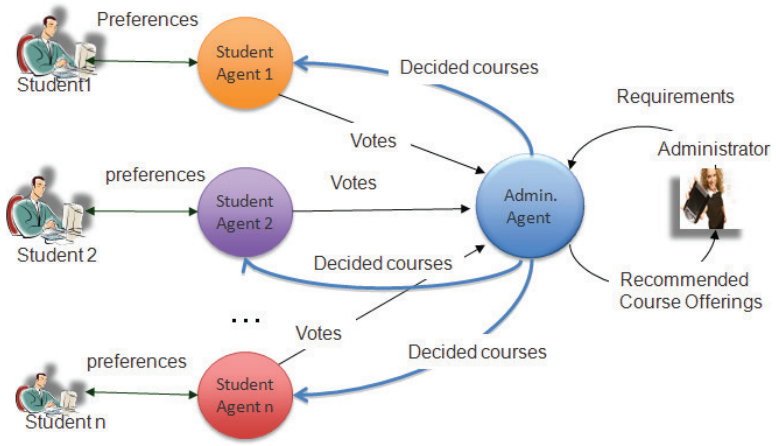


Fig. 1 The system diagram

When a student enters a program of study, the system creates a SA in the server. The SA runs in the background until the student is graduated from the program. The SA can be configured by the student with the basic information and profile of the student via a web-based interface. The process is started by the student who requires a study plan when s/he log into the system at the first time. Here we assume that the student has already chosen his/her program, and therefore, the curriculum of the program with a specific entrance year has been automatically decided by the system (Lin & Chen, 2013). This means that there are some mandatory courses the student must complete them to be graduated.

The COD problem is now concerned with the optimal offerings of courses for the upcoming semesters to meet the needs of the students in the program within budget and with scarce departmental resource, and maximize the course enrollment, while the goal for the students is to minimize the waiting time for desired courses to complete his/her graduation. In the system, human participants include students and program administrators, who exchange information and proposals in order to solve course selection tasks, and program planning tasks of the students, and the COD tasks of the program administrators. As shown in Figure 1, a society of software agents, including AD and a group of SAs work together to find an optimal solution

for COD. In this context, the students and the program administrator can be identified as the coordination entities, which are modeled as autonomous agents, where the SAs try to acquire the desired courses to take for some semesters from the AD agent. The primary responsibilities of a SA are of two folds: first, eliciting and reasoning about and learning the planning requirements and program preferences from students and generating personal plan; second, cooperating with AD agent to iteratively generate course delivery schedules collecting course selection preferences from students. On the other hand, the responsibility of an AD is conducting COD with SAs to generate course delivery schedules. The process of performing the responsibilities of SAs is detailed in the following two subsections:

3.1 Planning Program Preferences of SAs

For the planning of program preferences for SAs, a similar method as used in (Linden, Hanks, & Lesh, 1997) has been adopted. Let $\mathbf{C} = \{c_1, \dots, c_m\}$ be a set of *m* elective courses in a given program of study. Let $\mathbf{P} = \{p_1, \dots, p_n\}$ be a set of *n* attributes of the courses. For example, three attributes about assessment style are $p_1 = "Exam"$, $p_2 = "Project"$, and $p_3 = "Assignment"$. For a course c_i, an attribute value p_j is defined as a real number in $[0, 1]$, which represents the degree of correlation between c_i and p_j, $\{r_{ij}: i = 1, \dots, m; j = 1, \dots, n\}$. Therefore, course metadata for $\mathbf{C}$ and $\mathbf{P}$ forms a relation matrix, denoted as $\mathbf{R}(\mathbf{C}, \mathbf{P})$ representing domain knowledge between a set of courses $\mathbf{C}$ and a set of attributes $\mathbf{P}$. $\mathbf{R}(\mathbf{C}, \mathbf{P})$ can be acquired course-by-course from the course syllabi or from course instructors. For example, metadata for three courses, $\mathbf{C} = \{c_1, c_2, c_3\}$, and three attributes for each of them, $\mathbf{P} = \{p_1, p_2, p_3\}$, are presented as follows:

$$\mathbf{R}(\mathbf{C}, \mathbf{P}) = \begin{pmatrix} \mathbf{R}(c_1, \mathbf{P}) \\ \mathbf{R}(c_2, \mathbf{P}) \\ \mathbf{R}(c_3, \mathbf{P}) \end{pmatrix} = \begin{pmatrix} 0.3 & 0.5 & 0.2 \\ 0.5 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \quad (1)$$

From the relational matrix $\mathbf{R}$, it can be observed that a student preference for an attribute $p_j$with respect to a course c_i is expressed as a weight,$w_{ij} \in [0,1]$, indicating the degree of preference for c_i to p_j. The larger the weight is, the more preferable the student thinks. The weight is updated by a machine-learning algorithm, which is detailed in the next sub-section. Thus, for each student, *s*, it elicits a weight vector, $\mathbf{w}_p(s) = \{w_j: j = 1, \dots, n\}$. For example, for student "John", $\mathbf{w}_p(John) = \{1.0, 0.5, 0.1\}$ indicates that "John" prefers "*Exam*" to "*Project*", and the "*Assignment*" is less preferable.

To ranking a list of courses, a metric called Degree of Desirability (*DoD*) is used. For a course c_i, *DoD* for a student *s* is calculated by summing up the normalized product of the student's preference weights $\mathbf{w}_p(s)$ and course metadata $R(c_i, P)$, and then dividing by the sum of the student's preference item weights for a weighted sum. By the assumption that the preference structure is additive independence, we construct an error function which provides a partial ordering over all solutions. For course c_i, its preference weight, $DoD(s, c_i)$ is determined by the formula:

$$DoD(s, c_i) = \frac{1}{\sum_{j=1}^{n} \mathbf{w}_{pj}} \left(\mathbf{w}_p(s).R(c_i, P) \right) \quad (2)$$

For example, in the above example, student "*John*" specified his preference $\mathbf{w}_p(John) = \{1.0, 0.5, 0.1\}$, we have, $DoD(John, c_1) = 0.36$, $DoD(John, c_2) = 0.39$, $DoD(John, c_3) = 0.25$. Thus, if only considering John's preferences about the assessment style, the most preferable course is c_2.

3.2 Preference Weight Update Using Machine Learning

For updating preference weights, machine learning technique has been used. For the students start at an initial value for all preference attributes, which we set 1. In general, we want to have such a modification setup that the preference weights increase/decrease sharply with initial changes, but more slowly with the similar changes later on. This assures that trends change rapidly enough so that there will be significant differences in plans generated in the early phases of machine learning, but that once stronger trends are created they will not be greatly offset by any false assumptions. For learning the model of student preference weights, $\mathbf{w}_p(s)$, we use a tangential model (equation 3) with a maximum (horizontal asymptote) at 5 and minimum at 0. This range was chosen because the advisors who set the weights for the importance of preferences to individual courses are given the same range.

$$y = \frac{5}{3} \times \tan^{-1} \frac{x}{2} + 2.5 \qquad (where\ x \in \mathbb{Z}) \quad (3)$$

In Figure 2, the *y* axis is the weight and the weight modification process moves the position linearly along the *x* axis. It can be observed from this curve that when the weight is between 1.5 and 3.5, it is changed quickly and almost linearly with the changes in *x* positions, while at the higher and lower ranges changes are much slower. As mentioned before, the default starting value for all preference weights is $y = 1$, and so the starting *x* position on the weight curve is -2.52. During the preference elicitation stage, the

student can specify a preference attribute value as a level: primary, secondary, third, and forth. If a user selects a preference attribute as a primary preference, the x position for that preference is shifted right by 3, which brings the weight to nearly $y \cong 3\,(2.92)$. If it is a secondary preference, the x position for that preference is shifted right by 2, so its weight is around 2 (2.07). During the learning/training stage of preference weights, the weight y is updated with the plan selection and course selection from the selected plans.

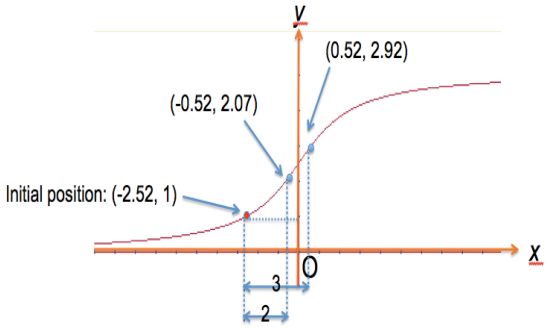


Figure 2: Proposed preference weight learning function

First, during the plan selection, the student is presented with several most preferable plans so that the student can choose a plan s/he prefers most. The interface will send the selection to the student agent who will determine the differences between the selected plan and the other plans. The agent will examine both plans and adjust preference weights depending on courses on the selected plan and the other plans. For a course c in the selected plan of student s, the weight $\mathbf{w}_p(s)$ for preference attribute p will be increased, if course-attribute relation $r(c_i, p_j)$ is non-zero. Similarly, for a course c_i in the not-selected plan of student s, the weight $\mathbf{w}_p(s)$ for preference attribute p_j will be decreased, if course-attribute relation $r(c_i, p_j)$ is non-zero. Second, after the student selected the most preferable (possibly not satisfactory) plan, the SA provides the students a way to approve or disapprove of courses individually (by buttons next to the course with "thumbs up" and "thumbs down" icons). Figure 3 shows a screenshot about this explicit feedback.

The preference weights are increased or decreased with respect to the x values along the weight curve by the following: $0.1 \times r(c_i, p_j)$ for the preference of the given course c_i and preference attribute p_j.

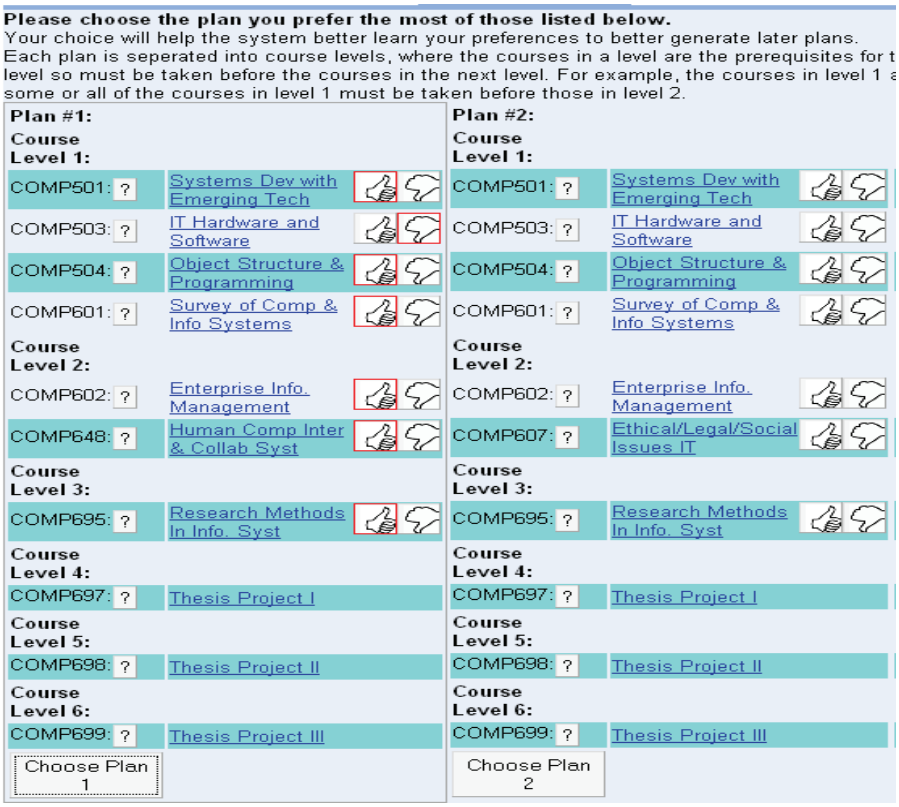


Figure 3: Plan selection and course selection interface

So, if a student s selects a plan with course A which has a preference, let us say the job objective "CTO," a course-attribute relation value of 5, the x position on the weight curve would shift right by 0.5 for the weight of that job objective in the student's preferences.

Similarly, for course B, if it has preference for the career track "Consulting" $r(B,"consulting")$ of 4, the x position of the weight curve would shift left by 0.4 for the weight of that career track in the student's preferences. This is done in two steps, first finding the x position for the old weight w_p' according to the formula,

$$x = 2 \times \tan\left(\frac{3}{5} \times \left(w_p' - 2.5\right)\right) \tag{4}$$

and then, the new weight w_p'' is calculated using the formula below.

$$w_p'' = \frac{3}{5} \times \tan^{-1}\left(x + \frac{\Delta x}{2}\right) + 2.5 \tag{5}$$

So, if the student has preference for "CTO" a weight of 3, the old x position would be determined to be 0.61 through using (4). From that a new weight would be calculated to be 3.34 according to (5). Similarly, if the student's weight for "Consulting" started at 2, it would decrease to 1.71.

4. PROTOTYPE IMPLEMENTATION

For the evaluation and benchmark of the proposed COD system, a prototype multi-agent system was implemented using JADE (http://jade.tilab.com). To test and evaluate the prototype under real-world conditions, a simulated environment was built, which allows simulating different scenarios by varying several parameters, such as the number of courses in the program, the divergence of preferences for course selection, and the must-offer courses in emergence cases. Figure 4 shows a screen shot of entering students' course selection preferences.

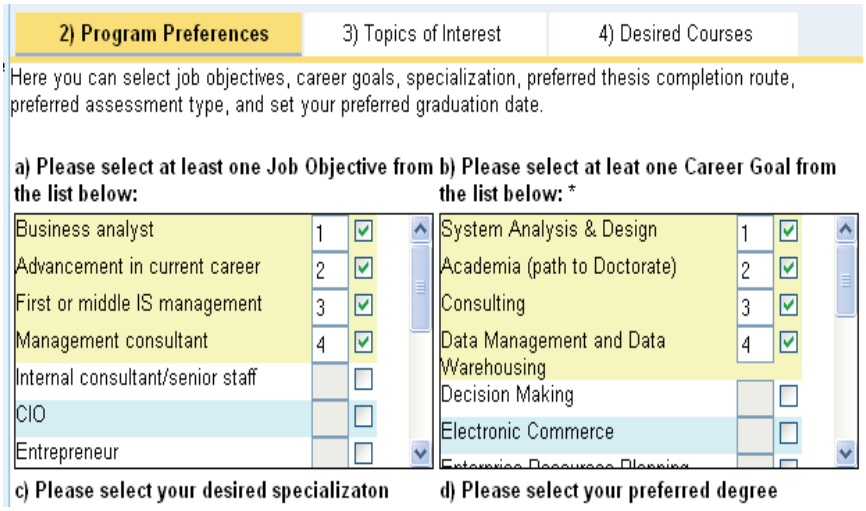


Figure 4: Program planning preferences

Table 1 shows simulated students' course selection preferences participated in the election process, where among 22 students 9 are freshman students f_1~f_9, 7 are sophomore or junior students s_1~s_6, and 6 are senior students e_1~e_6. There are 24 courses in the program. The simulated course Hits are shown in column H1 of Table 2. Utility for required courses is \$11,500. The candidate courses are denoted by C_1. By considering the curriculum and priority of some courses, the administrator determines a set of courses denoted by C_0 that must be offered in the next term to meet the requirements. The administrator also prepares an exclusion list containing

Table 1: Simulated students' course selection preferences

S	Course Priority	Course Grouping	Course sequences
f_1	c_{503} c_{601}	(c_{607} c_{610} c_{689})	{(c_{503} c_{601}), (c_{607} c_{610}), (c_{689} c_{667})}
f_2	c_{501} c_{503} c_{691} c_{504}	(c_{501} c_{503}), (c_{601} c_{504})	{(c_{501} c_{503}), (c_{504} c_{691}), (c_{648} c_{602})}, {(c_{602} c_{604}), (c_{648} c_{607})}
f_3	c_{501} c_{503} c_{691} c_{601}	(c_{501} c_{503}), (c_{689} c_{667})	{(c_{501} c_{503}), (c_{601} c_{691}), (c_{602} c_{689} c_{667})}
f_4	c_{503} c_{504} c_{601}	(c_{503} c_{504} c_{601})	{(c_{503} c_{504} c_{601}), (c_{604} c_{607} c_{648}), (c_{689})}
f_5	c_{601} c_{501} c_{504} c_{691}	(c_{501} c_{504})	{(c_{501} c_{504}),(c_{601}),(c_{691})}
f_6	c_{501} c_{503} c_{601} c_{691}	(c_{503} c_{501} c_{601} c_{691})	{}
f_7	c_{501} c_{504} c_{601}	(c_{504} c_{501}), (c_{602} c_{607})	{}
f_8	c_{501} c_{601} c_{691}		{(c_{501}), (c_{601}), (c_{602})}
f_9	c_{503} c_{504} c_{601}	(c_{504} c_{503}), (c_{607} c_{603})	{(c_{601}), (c_{602}), (c_{603} c_{607})}, {(c_{674}), (c_{695}), (c_{696})}
s_1	c_{602} c_{607} c_{691}	(c_{602} c_{691})	{(c_{602} c_{607}), (c_{695})}
s_2	c_{601} c_{691}		{(c_{601}), (c_{691} c_{602}), (c_{689})}
s_3	c_{695} c_{691}		{(c_{695}), (c_{617})}
s_4	c_{602} c_{691}		{(c_{602}), (c_{691} c_{695})}
s_5	c_{610} c_{691}	(c_{610} c_{691})	{}
s_6	c_{667} c_{607} c_{691}		{(c_{667} c_{691}), (c_{607})}
s_7	c_{605}	(c_{504} c_{503}), (c_{602} c_{604})	{(c_{602}), (c_{603}), (c_{604})}, {(c_{605} c_{607}), (c_{636}),(c_{667})}
e_1	c_{691} c_{695}		{(c_{695}), (c_{696})}
e_2	c_{695}		{(c_{695}), (c_{674} c_{696})}
e_3	c_{695} c_{691} c_{636}	(c_{691} c_{695})	{(c_{691} c_{695}), (c_{636})}
e_4	c_{695}	(c_{695})	{}
e_5	c_{695} c_{660} c_{691}	(c_{695})	{(c_{660}), (c_{695})}
e_6	c_{695} c_{691}	(c_{691} c_{695})	{(c_{691} c_{695}), (c_{617}), (c_{696})}

courses denoted by C_{-1} that absolutely may not be offered (perhaps because the responsible professor is on sabbatical or the course has a requirement that is only being offered at certain times of a year).

$$C_l = C_0 / C_{-1} \tag{6}$$

There are 88 participating students for the COD of the coming semester. Negotiation in *Round 1* assumes that $C_0 = \{c_{501}, c_{503}, c_{504}, c_{601}, c_{695}\}$; $C_{-1} = \{c_{602}, c_{604}, c_{617}, c_{636}\ c_{637}, c_{682}\}$; and $C_1 = \{c_{501}, c_{503}, c_{504}, c_{601}, c_{602}, c_{605}, c_{607}, c_{610}, c_{648}, c_{660}, c_{667}, c_{689}, c_{691}, c_{695}\}$. For U_{AD}, the actual cost to be paid for the courses offering is calculated with the following formula:

$$U_{AD} = Cost(C) - C_{ideal} = (b \times num_c + r \times \sum_{i=1}^{num_c} h_i) - C_{ideal} \tag{7}$$

where b is the base salary for one course to be paid to the instructor (e.g.,

b=\$5,000); r is the amount of money to gain $course\ fee - payment\ to\ the\ instructor$ for one course registration (e.g. r = \$500); h_i is the number of the registrations of course c_i (i = 1, 2, …, num_c), and C_{ideal} = \$75,000.

In negotiation round R2, c_{691} is chosen with a voting result of 60.77 getting Course Hits as shown in H2 in Table 2. For example, 64 people are going to take c_{691} since it is the first course chosen, and many students have it as their first pick due to it being on many plans and not having many prerequisites. Utility for the current course offering is \$15000.0. Comparing C_{ideal} (\$75,000), the administrator is still not satisfied with the required courses. Students are also unsatisfied with course offering as students' weight is 10.31.

Table 2: Course hits and votes in the simulated environment

	R_1	R_2		R_3		R_4		R_5	
	H_1	V_2	H_2	V_3	H_3	V_4	H_4	V_5	H_5
c_{501}	**23**	0.2	12		9		9		9
c_{503}	**17**	0.2	14		13		13		13
c_{504}	**11**	0.2	9		8		8		8
c_{601}	**20**	0.2	15		13		11		11
c_{695}	**6**	0.5	6		6		5		5
c_{607}		0.2		0.2		0.2		1.2	
c_{605}		5.0		**12**	10		10		10
c_{602}		1.6		1.6		1.6		1.6	
c_{610}		0.7		0.7		0.7		1.7	
c_{648}		2.0		2.0		**2.0**	4		3
c_{691}		**60**	64	0.5	64		64		64
c_{660}		1.0		1.0		1.0		1.0	
c_{689}		1.0		1.0		1.0		**2.0**	11
c_{667}		1.7		2.0		**2.0**	12		12

Now we select $\Delta U_{SR} = 3$. In negotiation round R3, the required courses are set as six: c_{501}, c_{503}, c_{504}, c_{601}, c_{695}, and c_{691}, votes see column "V_3" in Table 2. c_{605} is chosen in this round with a result of 12.0 by getting Course Hits shown in Column H3. Utility for current course offering is U_{AD} = \$21500.0. Comparing C_{ideal} (\$75, 000), the AD agent is not satisfied with the required courses. SA is also unsatisfied with course offering as weight is 7.34.

In negotiation round R4, the required 7 courses: c_{501}, c_{503}, c_{504}, c_{601}, c_{695}, c_{691}, and c_{605}. Fractional votes are shown in "V4" in Table 2. Courses c_{648} and c_{667} are chosen in this round, with a fractional result of 2.0. Course hits are shown in column H4 of Table 2. U_{AD} for current course offering is \$38000. Comparing C_{ideal} (\$75,000), AD agent is not satisfied with the course-offering result. SA is also unsatisfied with course offering as weight

is 4.2.

In negotiation round R5, the required 9 courses: c_{501}, c_{503}, c_{504}, c_{601}, c_{695}, c_{691}, c_{605}, c_{648}, and c_{667}. The fractional votes for this round are shown in V5 of Table 2. c_{689} is chosen in this round with a result of 2.0. U_{AD} = \$48000.0. So, AD agent is still not satisfied with the course offering. SA satisfied with the course offering as the SA weight is 0.86. Negotiation concluded with a course offering list: C_1 = {c_{501}, c_{503}, c_{504}, c_{601}, c_{695}, c_{691}, c_{605}, c_{648}, and c_{667}}. It can be seen from this example that H_1 = 77, H_2 = 110, H_3 = 123, H_4 = 136, H_5 = 145. As the num_c increases, the total course registrations increase accordingly. Here, we do not consider the size limit of a class. The negotiation and voting processes are described in details in (Lin & Chen, 2013).

5. CONCLUSIONS

In this paper, we presented the modeling and implementation details of student agent within a multi-agent based framework (Lin & Chen, 2013) for course-offering determination. It includes modeling of agent goals and behaviors and the protocols of agent-based coordination in dynamic decision making of individual students and the group decision-making of program administrators. One of the contributions of this work is the identification of a novel problem domain of determining a group of courses offering in educational institutions. The work can stimulate discussion of alternative points of view for how to solve the problem and lead to further discoveries. The second contribution of this work is the proposed architecture of multi-agent system that includes the algorithm to incorporate reasoning capabilities in the student agent, preference elicitation, and inference algorithm. Finally, there is a significant value to this mechanism design, where the administrative agent gets the ability to recommend suitable courses offering to departments in academic terms coordinating multiple student preferences, the course budget of the department, and the derived cost of the courses offered. Each student's preferences are translated into fractional votes that inform a negotiation process bounded by academic and resource constraints. The future work will focus on testing and deploying the system to turn it into a practical application. We will also explore the multi-winner election problem with exogenous constraints in other application domains.

REFERENCES

Brams, S., & Fishburn, P. (2002). *Voting procedure.* Handbook of Social Chice and Welfare .

Bruns, A. (2006). *Towards produsage: futures for user-led content production.* Murdoch University.

Burke. (2002). Hybrid recommender systems: survey and experiments. *User Modeling and Adapted Interaction, 12*(4), 331-370.

Conitze, V. (2010). Making decisions based on the preferences of multiple agents. *Communications of the ACM , 53*(3), 84-94.

Endriss, U. (2006). Monotonic concession protocols for multilateral negotiation. *International Joint Conference on Autonomous Agents and Multiagent Systems* . NY, USA.

Farzan, R., & Brusilovsky, R. (2006). Social navigation support in a course recommendation system. *International Conference on Adaptive Hypermedia and Adaptive Web-based Systems.* Dublin, Ireland.

Garcia, E., Romera, C. V., & Castro, C. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modeling and User-Adapted Interaction, 19*, 99-132.

Grupe, F. (2002). Student advisement: applying a web-based expert system to the selection of an academic major. *Academic Journal, 36*(4), 573.

Hamdi, M. (2006). MASACAD: a multiagent-based approach to information customization. *IEEE Intelligent Systems, 21*(1), 60-67.

Kannan, S., & Kasmuri, E. (2005). Empowering education managers in schools via a multi-agent system. *Malaysian Online Journal of Instructional Technology*, 17-24.

Kirn, S., Herzog, O., Lockemann, O., & Spaniol, O. (2006). *Multiagent engineering: theory and applications in enterprises.* Springer.

Krulwich, B. (1997). Lifestyle finder: intelligent user profile using large-scale demographic data. *MI Magazine, 18*(2), 37-45.

Lin, F., & Chen, W. (2013). Designing a multiagent system for course-offering determination. *Principles and Practice of Multi-Agent Systems.*

Lin, F., Newcomp, A., & Armstrong, A. (2012). A MAS approach to course offering determination. *IEEE International Conference on Web Intelligence and Intelligent Agent Technology.* Macau, China.

Linden, G., Hanks, S., & Lesh, N. (1997). Interactive assessment of user preference models: the automated travel assistant. *In the Proceedings of User Modeling.* Alberta, Canada.

Opera, M. (2007). MAS_UP_UCT: a multi-agent system for university course timetable scheduling. *International Journals of Computers, Communications and Control, II*(1), 94-102.

Resnick, P., & Varian, H. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Schafer, J. B., D. Frankowski, J. Herlocker, & S. Sen,. *Mehtods and Strategies of Web Personalization in the Adaptive Web*, (pp. 291-324).

Shen, W. (2002). Distributed manufacturing scheduling using intelligent agents. *IEEE Intelligent Systems, 17*(1), 88-94.

Smith, R. G. (1980). The contract-net protocol: high-level communication and control in a distributed problem solver. *IEEE Transactions on Computers, C-29*(12), 1104-1113.

Smyth, B., & Cotter, P. (2001). Personalized electronic program guides for digital TV. *AI Magazine, 22*(2), 89-98.

Smyth, B., Shang, Y., Shi, H., Chen, & Shing, S. (2001). An intelligent distributed environment for active learning. *ACM Journal of Educational Resources in Computing, 22*(2), 89-98.

Tang, T., & McCalla, G. (2005). Smart recommendation for an evolving e-learning system: architecture and experiment. *International Journal of E-Learning, 4*(1), 105-129.

Vainio, A., & Salmenjoki, K. (2005). Improving study planning with an agent-based system. *Informatica, 29*, 453-459.

Weiss, G. (1999). *Multiagent systems, a modern approach to distributed artificial intelligence.* MA, USA: MIT Press.

Winikoff, M., & Padgham, L. (2013). *Agent oriented software engineering*. MA, USA: MIT Press.

Wooldridge , M., Jennings, N., & Kinny, D. (2000). The Gaia Methodology for Agent-Oriented Analysis and Design. *Journal of Autonomous Agents and Multi-Agent Systems, 3*, 385-312.

ACKNOWLEDGEMENTS

We thank Natural Sciences and Engineering Research Council (NSERC), Canada, and Research Incentive Grant (RIG), Athabasca University, Canada for their support to this research project.

Chapter 10

MULTI-AGENT WELL SCHEDULING: A PROTOTYPE IMPLEMENTATION USING CNP AND JADE

Jivko Hristov[1], Graham Lange[2], Fuhua Lin[1], M. Ali Akber Dewan[1], Xiaokun Zhang[1], Saadat Khan[1]
[1]*School of Computing and Information Systems, Athabasca University*
[2]*Encana Corporation, Canada*

Abstract: Efficient task allocation and resource scheduling have been challenging in oil and gas industries for many years, because they are influenced by a number of factors including resource availabilities, environment, regulations, stakeholders, finances, and market. With the advancement of information communication technologies (ICT), the oil and gas industries get the opportunities to increase production and to minimize operation costs through efficient resource management and task scheduling. This paper presents a prototype implementation of daily well scheduling using Java Agent Development Framework (JADE) — a multi-agent system platform. Coordination mechanism among the agents are implemented using the traditional Contract Net Protocol (CNP) which enables a flexible and efficient resource allocation leading to an intelligent management of the available resources and dynamic scheduling of the tasks across the well lifecycle. In the prototype model, sequence diagrams and class diagrams are used to show coordination mechanisms between different agents. Three different use cases initially demonstrate its effectiveness of the CNP-based coordination of multi-agent system approach to well scheduling. Future work on well scheduling in distributed and online environment is discussed.

Key words: Multi-agent systems, well scheduling, Contract Net protocol, JADE, business process modeling.

1. INTRODUCTION

Well scheduling is a highly dynamic in nature and complex problem in the oil and gas industries. Large oil and gas industries plan for their active drilling season in advance to utilize their equipment (or available resources) to the maximum and achieve high return on investments. Unfortunately, the active drilling season has limited window opportunity to complete all business goals. The work activities of the oil and gas industries largely depend on many internal and external factors, such availability of the resources, weather, regulations, and health and safety inspections. These factors are hard to manage by planners because it involves hundreds of processes and dependency tasks which change dynamically. This dynamic nature of the working environment warrants flexible technological solutions that will be able to allocate tasks and schedule resources intelligently to maximize industrial benefits.

A number of methods for well scheduling have been proposed in the last few years. To schedule the well activities, Hasle et al. (1995) developed a model by defining a particular well activities problem and generating a high quality and feasible schedule that can be inspected and modified by the user through interactive Gantt visualizers. In another work, Dimitrios et al. (2009) proposed a mixed integer nonlinear programming model. However, these models either require manual interaction or face challenges with the dynamic nature of the well scheduling. One possible solution would be the multi-agent systems, which can allow dynamic allocation of the tasks and schedule of the resources through agent negotiations, while managing all the related internal or external factors in a flexible manner. Multi-agent systems technologies have been widely used in complicated systems, which play a role in solving distributed and complex problems coordinately (Wooldridge & Jennings, 1995). In the multi-agent systems, agents can be software entities, computer programs or distinct objects in a larger software model that act autonomously on behalf of their users (Weiss, 1999; He et al., 2008). To resolve the dynamic scheduling problem we not only need agents but also need a way of effective communication between the agents and a way to negotiate terms and conditions to process the work.

Lange and Lin designed a system for well scheduling based on multi-agent systems platform (Lange & Lin, 2014). This system design is capable of negotiating among agents with agility and flexibility for a better solution of well scheduling. A prototype design and implementation of the above system using Java Agent Development Framework (JADE) and Contract Net Protocol (CNP) have been detailed in this paper. More specifically, “Add a well” among many different scenarios has been described. This scenario adds wells to the system and establishes communication to the service

providers who can provide their bids. The tasks are assigned to the service providers who provide earliest completion time with minimal cost proposal. Percept sources, such as weather, roads condition, and health and safety have been considered. Three use cases are explained within the prototype implementation for performance analysis. The advantages of the proposed multi-agent system in "Add a well" scenario have been identified.

2. SYSTEM ARCHITECTURE

2.1 Overview

The well scheduling is a very complex and dynamic process with various tasks involved in commencing of a well. At a higher level, the tasks identified by the researchers are divided into five distinct phases: *Landman*, *Construction*, *Drilling*, *Completion*, and *Facilities*. The tasks are tightly coupled and can only be executed in sequential order. For example, the *Construction* task cannot be started before the *Landman* task is completed for a given well, or *Drilling* task cannot be begun before the *Construction* task is completed. The scheduling module needs to take this precedence information into account and expect to schedule all the tasks using an optimal plan. Each high level task is comprised of many lower level sub-tasks which must be completed before the overall process can move forward to the next milestone. A high level system diagram of the proposed business process model for well scheduling is shown in Figure 1, where several actors (or agents) are participating in the process of allocating tasks and scheduling resources for an optimal scheduling solution. It can also be observed that there is an actor representing each phase of the project like *Landman*, *Construction*, *Drilling*, *Completion*, and *Facilities*.

The well scheduling is comprised of many different scenarios such as "Add a well", "Remove a well", "Facilities Complete", "Competition Finished", "Drill Completion", "Construction Complete" "Stop Work", etc., each of which has many different functionalities. In this paper, the implementation of "Add Well" scenario with its functionalities has been described. The system determines the order of tasks for "Add Well" scenario as per predefined plan and the best scheduling time when the work can be completed. The scheduling decision is made based on previous service commitments thus the service providers will try to provide best available services in time. However, the work can be disrupted due to many external factors which stimulate a simple stop order. After completing the scheduling,

the system is presented with the scheduled tasks which include the cost for every part of the whole project.

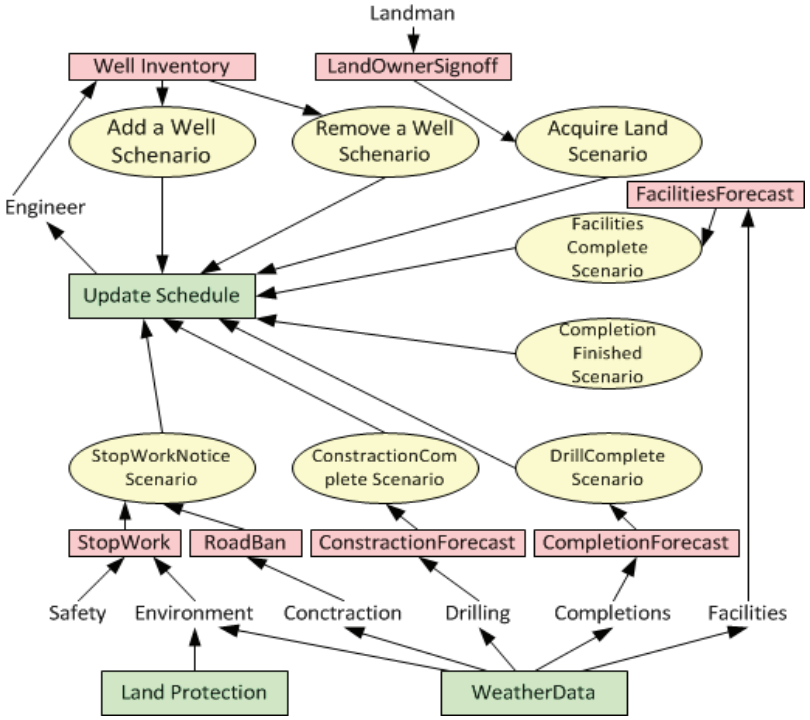


Figure 1: System architecture for multi-agent based well scheduling (Lange & Lin, 2014)

A high level sequence diagram is shown in Figure 2, which defines the high level message flow between agents. The process is initiated by the Engineer agent who adds a well to the system to be scheduled. The message is sent to the well agent who starts the negotiation process with the scheduling agents. There are five scheduling agents one for each high level phase – *Landman*, *Construction*, *Drilling*, *Completion* and *Facilities*. Each scheduling agent communicates with its respective service providers to find the best service offer which meets the shortest execution time utility. The utility function implemented by any of the agents can be developed to meet any business requirement. For this prototype the utility function will try to find the shortest path to complete the given work. Based on the high level system design, this paper defines the agents below with their respective functionalities. A high level sequence diagram of agents' interactions is shown in Figure 2.

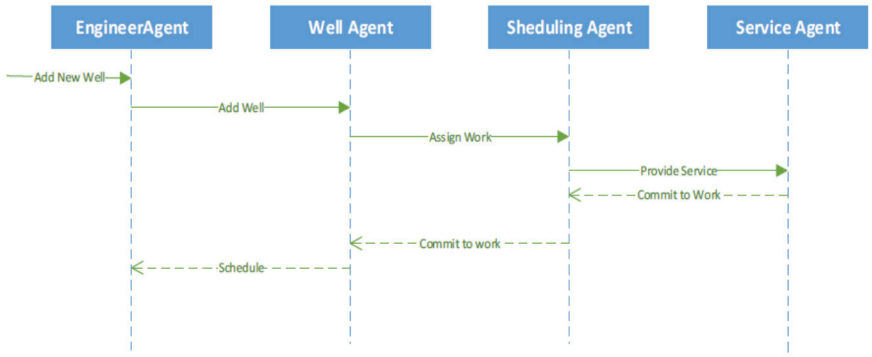


Figure 2: Sequence diagram of the multi-agent-based well scheduling system

2.2 Engineer Agent

This agent initiates the process of adding wells to the system to be scheduled. In the prototype, this functionality is simulated by implementing timer behavior which adds a new well in every 20 seconds to the maximum of nine wells. The upper limit was put in place to study the use cases defined in the research paper (Lange & Lin, 2014). If the limit is removed, the system will have no upper limit and continue to add wells after every 20 seconds for scheduling. The engineer agent executes three scenarios and for each scenario it adds three wells.

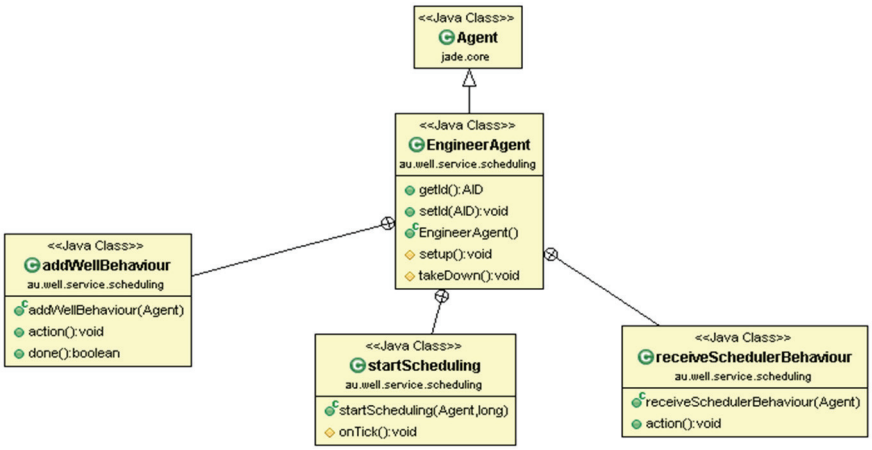


Figure 3: Class diagram for the engineer agent

The purpose of this design is to simulate the three different scheduling options the intelligent agents can use when scheduling the tasks. This

scenario follows the close guidelines of the design in Lange and Lin's paper (Lange & Lin, 2014). Messages exchange by the agents is designed with the maximum flexibility to allow for future expansion of this application. Every time the agent wakes up it initiates the addWellBehaviour which send message to the well agent to add a new well to be scheduled. Class diagram of the engineer agent is shown in Figure 3.

2.3 Well Agent

The role of this agent is to accept requests from the engineer agent when a "Add a well" is initiated, to request for scheduling various tasks to the scheduling agent, and to send back the overall schedule and related cost to the engineer agent. In this design, the well agent plays the role of coordinator of the scheduling works. The agent extends JADE agent framework and implements several private classes to support the functionality as per well agent class diagram as shown in Figure 4.

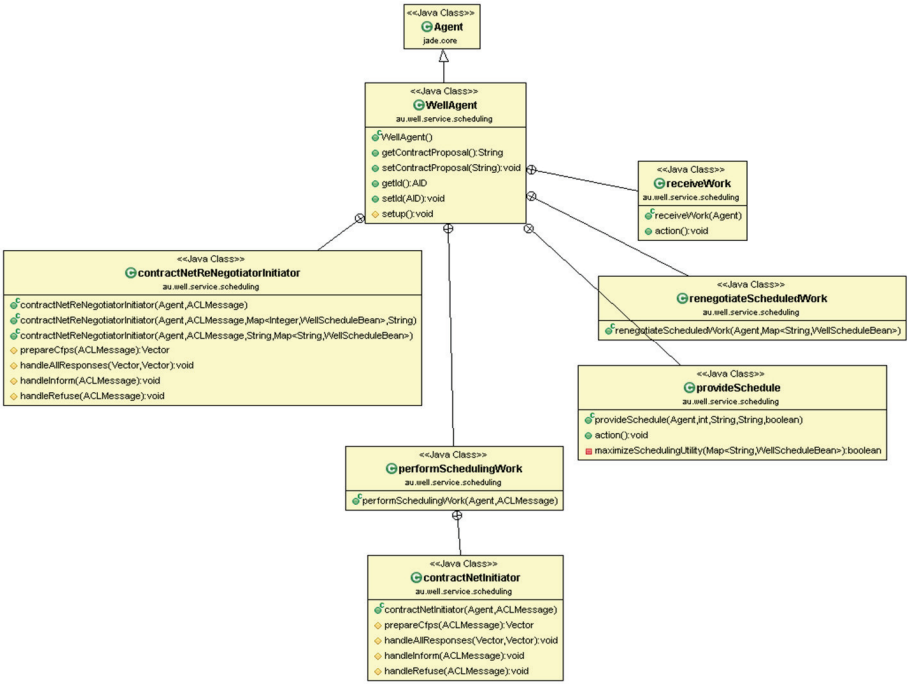


Figure 4: Well agent class diagram

Once the scheduling request is received from the engineer agent, the well agent creates a sequential plan using SequentialBehaviour which defines the order of execution of scheduling work requests to the five scheduling agents. The sequential plan helps to control the sequence of execution of requesting

schedules from the various scheduling agents when a new well is added to the system. Sub-behaviours are added to the plan by means of using addSubBehaviour() method which ensures that tasks are executed in the order they were added to the execution plan. The sequential behaviour can be instructed to terminate execution after the first child completes or after the last child finishes, it all depends on the business requirements. As previously described the order of scheduling work is sequential and the sequential behaviour helps enforce this rule and fire scheduling requests in the predefined order of execution (*Landman, Construction, Drilling, Completion and Facilities*).

Since the basic business requirements are that the schedules should not be overlapped, it implies that scheduling agents must share at the very minimum completion date with the next agent in the execution chain. Given that the well agent plays a coordinator role the well agent receives the schedule from every scheduling agent before it fires the request to the next one in the execution list which is controlled by the sequential behavior. The well agent uses the CNP to negotiate schedules with the respective scheduling agents. When the well work is fully scheduled the well agent uses a utility function to determine if the schedule of the works can be optimized. The optimization decision is based on a stop work order introduced in the schedule. The stop work order causes the well work schedule to be extended by the duration of the stop order thus costing the company additional cost.

The well agent ensures that the schedules returned to the engineer agent are optimal; therefore it requires the scheduling agents to re-plan the order of execution and find optimal work schedules across different wells (see sequence diagram in Figure 2). To support this functionality the solution introduces two new system operations – *Change* and *Move*. This new operations will request the scheduling agents to prioritize their work based on new provided start dates across the work they have already committed to. Alternatively more sophisticated implementation of this process could use iterated CNP which will be discussed further down this paper. The agent implements the business logic using different plans which drive out the decision of well scheduling.

The well agent recognizes this event by reviewing proposed schedules before returning them to the engineer agent. When reviewing schedules the well agent verifies schedules not only for the current well which is being created by it goes back to the previous schedules to find potential work stop gaps which can be utilized to optimize current and previously committed schedules to reduce work duration. The scheduling agent will accept the optimization request and assess the available options.

2.4 Scheduling Agents

The application design introduces a scheduling agent for every task in the process flow, for example one scheduling agent for *Landman*, another one for *Construction* etc., as per the scheduling agent class diagram as shown in Figure 5. Since the application has many scheduling agents, the design abstracts the common Contract Net protocol framework to the abstract class SchedulerAgent so every agent can implement the abstract methods and reuse the common functionality of the base class.

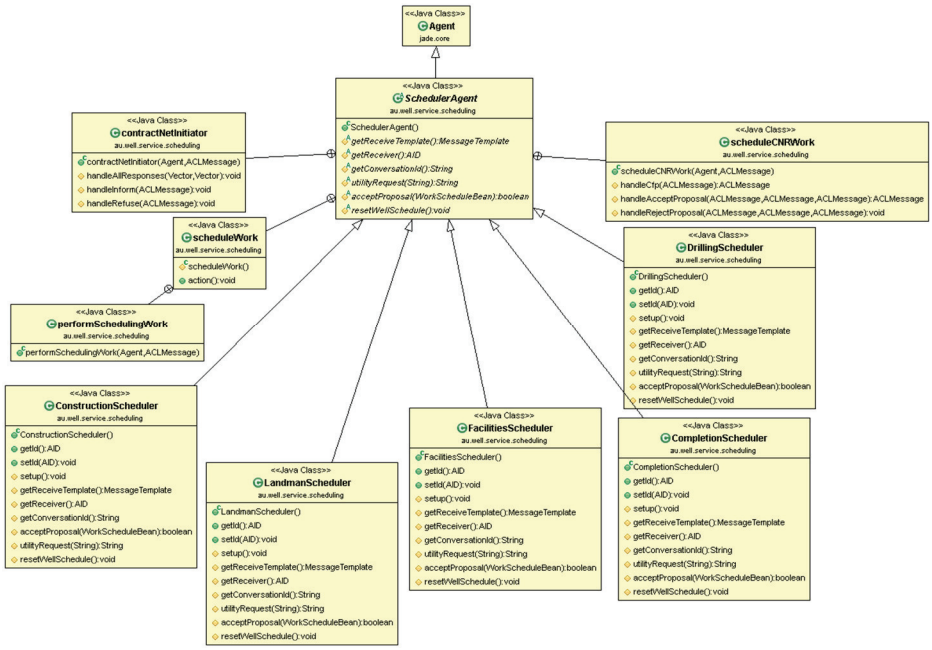


Figure 5: Scheduling agent class diagram

The base class will call each abstract method which will be implemented by all five scheduling agents. Each individual agent will know the services providers and maintain internal work schedule. The scheduling agent will support three operations like *AddWell*, *Change* and *Move*. The first operation, *AddWell*, will support adding new well to the schedule. Before any new work is taken on the scheduling agent executes its utility function to find best offer it needs to meet the requested demand. This proposal is sent to the service agent for commitment or counteroffer. As part of the contract net negotiation protocol when the scheduling agent receives a proposal from the servicing agent it validates all proposal to find the best offer. Each scheduling agent can contact one to many service providers with work request which they can bid on. When all bids are received as part of the

contract net negotiation process the scheduling agent reviews all offers and selects the best offer, the offer that is either matching the proposed start date or the one with shortest duration. The scheduling agent declines all offers with the exception of the one it accepts for which it requests the service agent to commit to it.

The second operation *Change* will indicate to the scheduling agent to consider rescheduling work for various wells based on the proposed start dates. The agent will validate its internal committed work and decide if optimizations are possible. If optimization is possible the agent will engage the service provider to arrange for new commitments. The scheduling agent always verifies the offers coming back from the service providers against their internal records. If the service provider offer is different from the recommended schedule which the scheduling agent provided upon submitting the request for work messages to all agents it either can decline it or accept the best offer. When the best offer is accepted and committed to by the service provider the scheduling agent records it in its internal memory. Irrespective of the operation executed the underlying negotiation protocol used is the same. All interactions between agents when negotiation is required use the CNP framework.

2.5 Service Agents

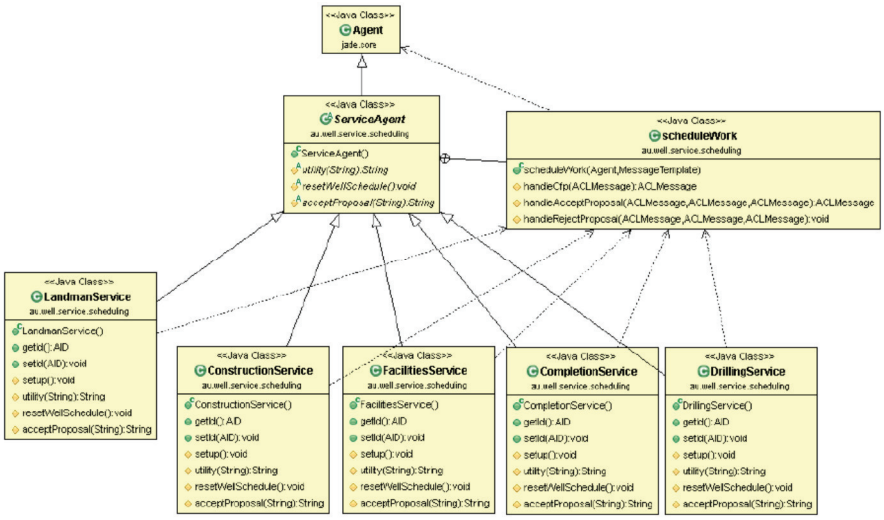


Figure 6: Service agent sequence diagram

This agent represents the service providers in the system. The service agents' base class (ServiceAgent) implements all common service functionality such

as the contract net responder and provides several abstract methods which are implemented by the service providers. Each service provider as per Services Class Diagram as shown in Figure 6 implements its own utility functions and corresponding abstract methods. Services agents bid on proposed work by the scheduling agents (see sequence diagram in Figure 2). The service agents only implement Contract Net responder functionality because the services only bid on work and either their bid is accepted or rejected.

3. PROTOTYPE IMPLEMENTATION IN JADE FRAMEWORK

This section describes the implementation details of the Contract Net protocol in JADE framework. The implemented prototype extensively uses the Contract Net protocol to exchange messages between agents which use the protocol to negotiate for best solutions by optimizing their internal goals.

In CNP for agent negotiation, the process (as Figure 7 shows) starts with contractor agent broadcasting works to service provider agents. The service provider agents send their proposals to the contractor agent. The contractor agent looks at all proposals and finds the best proposal that maximizes its utility function. Once the optimal proposal is found, the contractor agent sends accept message to the agent that provided the best offer and reject message to the other agents. The agent that receives the accept message needs to commit to the offer by sending a commit message to the contractor.

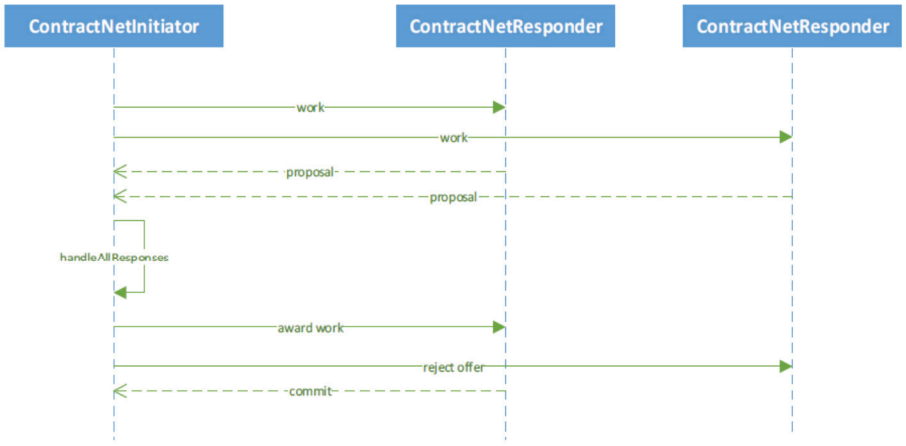


Figure 7: Sequence diagram of Contract Net protocol.

In the prototype implementation, the well agent is used as a coordinator which is responsible for providing requirements from one scheduler agent to the next aiming to achieve its goal of completing scheduled well work in the shortest time. The scheduling agent receives its work from the well agent with proposed start time of the work. Every scheduling agent tries to achieve this target by requesting the service providers match the proposed timelines. The benefit with this architecture is that the scheduling agents don't know of each other's existence. They are decoupled from one another and work independently. Another positive side for this architecture is that if additional work or new phase is introduced in the well work schedule, it can easily be added to the execution plan of the well agent and none of the scheduling agents will need to be concerned with it nor there will be any changes for those scheduling agents.

Alternatively this solution can be built with different architecture where well agent initiates the work by broad casting it to the agents. In this model the agents can be chained and each agent knows which other agent it can receive work from and once completed who to pass the work to. This model is obviously a bit more complicated and harder to manage due to fact that agents will need to have previous knowledge about their surrounding agents in the chain of work.

Another observation about the agent platform is that when agents fail they stop receiving messages and become unresponsive which may be an issue if this platform is used to implement enterprise solutions because the system needs to be stable and reliable. This fining will require the application to be well designed and developed to handle and deal with any unforeseen exceptions during application execution to avoid the agent to become unresponsive.

Scalability of the platform will need to be further assessed. The business logic in each method of the behavior is executed only after the previous method exits. This constraint will require very careful application design so that behaviors are executed efficiently. The behaviors should not be developed with blocking logic because if the business logic requires other work to be executed while waiting it will never be executed since the blocking behavior method will never exit to allow the scheduler to invoke the next behavior work in the pipeline. The reason for this constrain is because agents are designed to execute in environments will limited resources thus this architecture definitely suites systems in that space. However, in larger environments where this constrain is not applicable will see this requirement as a negative constrain.

Sharing data between different behaviors is more challenging than in regular applications. The reason for it is because behaviors are pre-

constructed before they are executed thus information that is obtained by the first behavior via some business functionality can't be passed to the next behavior in the chain of execution because it was not available when the plan was created. The way this problem is resolved in this environment is by either using local variables in the agent or its parents (Bellifemine et al., 2007). However, keeping data stored in the agents or parents variables is not a best practice because it prevents common logic to be reused. Therefore, it is best if data is externalized from the agents and stored in Data Store class included in the jade framework.

JADE platform executes logic in a single thread using the concept of behaviors. The behaviors provide several different implementations that the developers can use to solve various business requirements. This project utilizes several of the JADE behaviors to solve this assignment design requirements. The sequential behavior pattern is used to execute tasks in specific order to ensure that no one task is executed before its precedent. Well agent is perfect example where this behavior is used. As per the business requirements the first agent needs to receive the commitments of the first service provider before it can provide requirements to next service provider in the chain of execution. This pattern is heavily utilized to enforce sequential execution of tasks. Another behavior used is one shot behavior. This framework is used in event when specific task is required to be executed just once. Typical use case for using it is sending a message to another agent without requiring waiting for the response. Another interesting functionality implemented using behavior framework is ticker behavior. It gives the application capability to execute behaviors within preprogrammed intervals. Very important behavior called cyclical is used to constantly listen for arriving messages. Since the agents constantly exchange this behavior can help implement asynchronous message exchange pattern where agent can send a message to other agents and not necessarily wait for any response immediately. This behavior can help implement asynchronous message pattern exchange between agents.

4. CASE STUDIES

4.1 Use Case 1

This use case depicts the three well schedules in which there are no external forces that impact the committed work timelines by the various parties. As outlined each phase takes exactly the allowed time to complete before the next phase can begin. The test case implemented confirmed Lange and Lin (Lange & Lin, 2014) estimate that if the work is completed as per

requirements the scheduling of the three well test solution is optimal. The work duration for this small project completes in 12 months.

Tasks	**Well #1**	**Well #1**	**Well #1**
Land	$50,000	$50,000	$50,000
Construction	$100,000	$100,000	$100,000
Drill	$2,000,000	$2,000,000	$2,000,000
Complete	$1,000,000	$1,000,000	$1,000,000
Facilities	$500,000	$500,000	$500,000
Total	**$3,650,000**	**$4,650,000**	**$3,650,000**
			$10,950,000
	Total Time		**13 months**

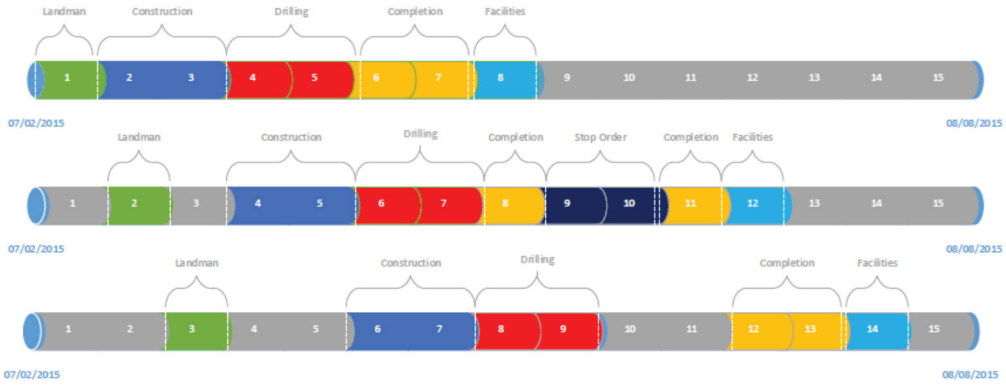


Figure 8: Scheduling result for Case 1

4.2 Use Case 2

Use case two implementations execute the same model where three wells will need to be scheduled. The difference here is that stop order is introduced in the middle of the completion phase for 2 months. Naturally this will push out the completion phase by two months resulting in delays for the overall schedule which not only impacts well two schedule but also well three schedule. Well three's schedule is also delayed because the completion work can't commence as planned. This stop order directly impacts the optimal schedule which results in 2 months delay. Therefore the entire duration of this project would be equivalent to 14 months.

Tasks	**Well #1**	**Well #1**	**Well #1**

Land	$50,000	$50,000	$50,000
Construction	$100,000	$100,000	$100,000
Drill	$2,000,000	$2.000,000	$2,000,000
Complete	$1,000,000	$1,300,000	$1,000,000
Standby		$400,000	
Facilities	$500,000	$500,000	$500,000
Total	**$3,650,000**	**$4,350,000**	**$500,000**
			$11,650,000
		Total Time	**17 months**

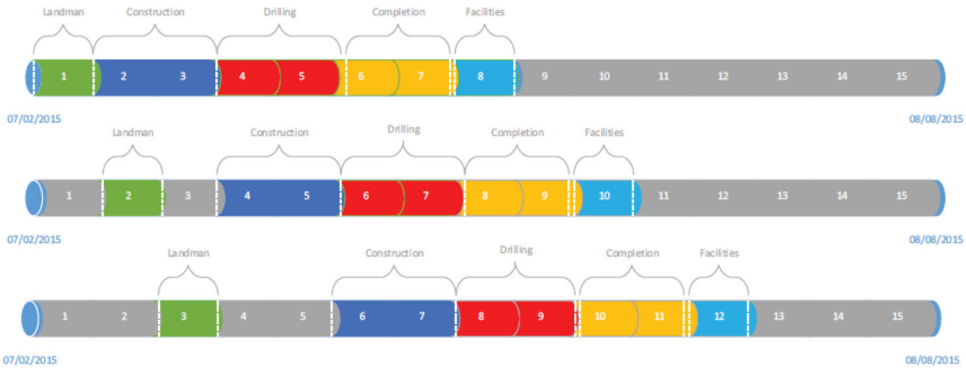


Figure 9: Scheduling result for Case 2

4.3 Use Case 3

In this use case the same stop order is introduced into the second well completion phase schedule however the agents consider optimizing the schedule when stop order is introduced in any phase of the well schedule. The optimization reduces the duration of the project by two months helping the agents achieve their overall goal.

Tasks	**Well #1**	**Well #2**	**Well #1**
Land	$50,000	$50,000	$50,000
Construction	$100,000	$100,000	$100,000
Drill	$2,000,000	$2,000,000	$2,000,000
Complete	$1,000,000	$1,300,000	$1,000,000
Standby		$200,000	
Facilities	$500,000	$500,000	$500,000
Total	**$3,650,000**	**$3,950,000**	**$3,650,000**
			$11,450,000

		Total Time	**15 months**

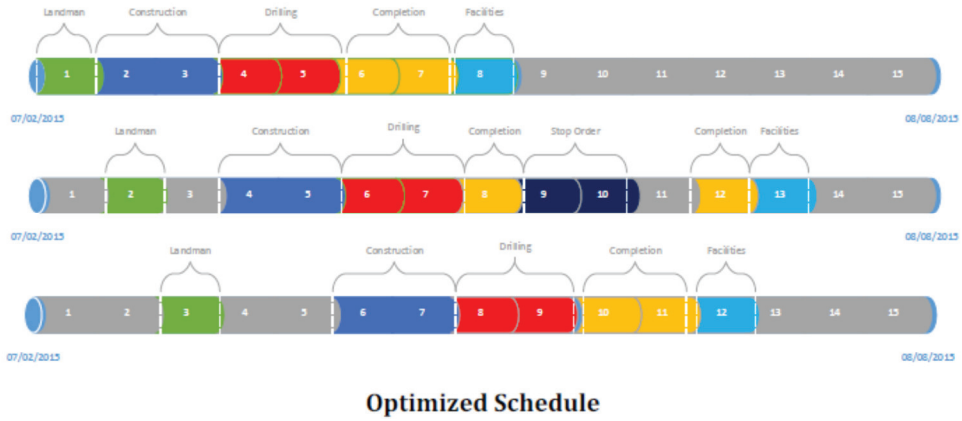


Figure 10: Scheduling result for Case 3

5. CONCLUSION

Multi-agent based well scheduling approach based on the notion of virtual enterprise in the oil and gas industry can solve a number of problems that conventional approaches cannot solve. By implementing a multi-agent system to well scheduling we can automate this manual effort. Multi-agent systems approaches are well suited for a dynamic environment and by working through the design it is evident that the approach will work. The multi-agent approaches allows for a certain amount of flexibility and timeliness not provided in traditional systems.

Coordination of CNP with different agents was used as the negotiation mechanism for solving the resource and scheduling problem. This should be a suitable approach to solving the problem of negotiating multiple resources across the schedule. However, the limitation of the current implementation is that it is one task version of CNP. There are weaknesses of this method: lack of optimality. This lack of optimality is due to decisions that lead to the myopic behavior of decision-making entities. For example, the case of a single task in reactive scheduling, but this is actually a degenerate case of the CNP. That is, the agents are ignorant in temporal aspect without taking into account what the other proposals that may arrive later. The temporal aspect comes from the fact that the bidders are aware only of the requests for bids already received, and not of those that are on the point of arriving. An agent

may commit it to carry out a task for which it is not really very well suited, and therefore miss contracts that would correspond better to its true qualities (van Parunak, 1987). Therefore, future research could include additional negotiation approaches such as Combinational Auctions (Sandholm, 2002).

What is required to further this work is to further implement the solution and run a large number of tests to ensure that the results are acceptable to validate and verify the model. Since the benefits of applying multi-agent approach to the well scheduling problem is directly proportional to the size of the overall well program, the tests should include a test for scalability. Tests should be conducted for 50,100, and 1000 well programs. The results should be compared to traditional systems currently in use. These systems include manual scheduling, Microsoft Project and Oracles Primavera. The comparisons would further support the implementation of a multi-agent approach in an enterprise environment. In addition to comparing the system to scheduling and planning tools, further research into comparison to a traditional BPM method would help to understand which is a better application. Further research on the coordination of multiple Contract Net protocols in a system should be conducted. This would help to determine whether this approach is the most applicable for the well scheduling problem.

REFERENCES

Arash Mousavi, A., Nordin, M., and Ali Othman, Z. (2011, April 19). Ontology-driven coordination model for multiagent-based mobile workforce brokering systems. *Applied Intelligence, 36*(4), 768-787. doi:10.1007/s10489-011-0294-z

Bellifemine, F., Caire, G., and Greenwood, D., (2007). Developing multi-agent systems with JADE, Wiley

Gerogiorgis, D., Kosmidis, V., and Pistikopoulos, E. (2009). Mixed Integer Optimization in Well Scheduling. *Encyclopedia of Optimization*, pp. 2247-2270. doi:10.1007/978-0-387-74759-0_395

Hasle, G., Haut, R., Johansen, B., & Ølberg, T. (1995). Well activity scheduling-an application of constraint reasoning. *Artificial Intelligence in the Petroleum Industry: Symbolic and Computational Applications II*, 209-228.

He, L., Liu, Y.-X., Xie , H.-l., and Zhang, Y. (2008). Job shop dynamic scheduling model based on multi-agent," in Control and Decision Conference, Yantai, Shandong, China. 829 - 833

Lange, G., & Lin, F. (2014). Modeling Well Scheduling as a Virtual Enterprise With Intelligent Agents. Proceedings of *IEEE*

Computational Science and Engineering Conference. Chengdu, China. 89-96

Parunak, H. V. D. (1987). Manufacturing Experience with the Contract Net. In Distributed Artificial Intelligence, M. Huhns (Ed.), Pitman.

Sandholm, T. (2002). Algorithm for optimal winner determination in combinatorial auctions. *Artif. Intell.* 135, 1-2, 1-54.

Weiss, G. (1999). *Multiagent Systems A Modern Approach to Distributed Modern Approach to Artificial Intelligence.* MIT Press.

Wooldridge, M., and Jennings, N. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review, 10*, 115-152.

ACKNOWLEDGEMENTS

Here we thank Natural Sciences and Engineering Research Council (NSERC) of Canada and Encana Services Inc. (Canada) for its financial support to the research project.

Author Index

Al-Shamali, F. 17
Bagheri, E. 105
Baldiris, S. 39
Bernard, J. 39
Betrie, G. D. 27
Boulanger, D. 57
Chang, M. 75
Chang, T.-W. 39
Deng, B. 27
Dewan, M. A. A. 121, 137
Graf, S. 39
Guillot, C. 93
Guillot, R. 57, 93
Hobbs, D. 39
Hristov, J. 137
Khan, S. 137
Kinshuk 57, 93
Krys, G. 105
Kumar, V. 57, 93
Kumar, V. S. 1
Kurcz, J. 39

Lange, G. 137
Lewenza, S. 1
Lin, F. 121, 137
Newcomb, A. 121
Reckseidler-Zenteno, S. 1
Seanosky, J. 57
Sodjahin, B. 1
Tortorella, R. A. W. 39
Wang, J. 1, 27
Zhang, X. 137

October 2015

Publisher: Athabasca University

http://fst.athabascau.ca/

ISBN 978-1-987973-00-6 (print)
ISBN 978-1-987973-01-3 (online)

