Mining Student Behavior Patterns in Reading Comprehension Tasks

Terry Peckham
Department of Computer Science
University of Saskatchewan
110 Science Place
Saskatoon, SK., Canada
tep578@mail.usask.ca

Gord McCalla
Department of Computer Science
University of Saskatchewan
110 Science Place
Saskatoon, SK., Canada
mccalla@cs.usask.ca

ABSTRACT

Reading comprehension is critical in life-long learning as well as in the workplace. In this paper, we describe how multidimensional k-means clustering combined with Bloom's Taxonomy can be used to determine positive and negative cognitive skill sets with respect to reading comprehension tasks. This information could be used to inform environments that support students improving their meta-cognitive skills.

Keywords

Data mining, k-means clustering, Bloom's taxonomy, reading comprehension, cognitive strategies.

1. INTRODUCTION

Anderson and Pearson [4] in their seminal work on reading comprehension describe three different cases where reading comprehension is a problem. First, a person having difficulty reading is likely to have gaps in knowledge. Prior knowledge is necessary in the determination of what he/she can currently comprehend. Second, the reader can have an incomplete understanding of the relationships that exist among facts on a certain topic. Since the current knowledge base is used to create all of the relationships on a topic, new arbitrary information can be a source of confusion, slow learning and slow processing, which leads to unsatisfactory reasoning. Third, readers are unlikely to be able to make correct inferences about the material in order to arrive at a coherent overall representation of the topic. The creation of a coherent representation for a topic requires the drawing of precise, integrated inferences. Often poor readers do not perform these tasks either routinely or spontaneously [8]. Any reading comprehension tools or models need to be able to address these problems with deep comprehension.

The reading strategy instruction method is one of the most often suggested methods for enhancing reading ability [18, 20]. This particular method deals with problems on the vocabulary and sentence levels [2], and on higher level issues such as text comprehension [2,14]. Other recommended approaches include determining the main message of the content (e.g. summarization), the use of textual enhancements (e.g. illustrations, mental images), question and answer drills (e.g. self-questioning) and practicing meta-cognition (e.g. through comprehension monitoring) [9]. However, the most successful reading strategies combine methods rather than one single technique [14].

There are several barriers to the adoption of multiple strategies within the classroom setting [14]. First, there is a large amount of training that is required for the teachers to become familiar with the strategies in order to employ them within the classroom setting. Second, there is a considerable time requirement for teachers to prepare the course materials. Third, getting the students to apply the strategies in daily life can be extremely complex. Therefore, the creation of environments that help relieve the teacher of some of these complexities would be of great benefit.

There are several learning environments that aid students with their reading comprehension. Some of the more prominent are Project Listen, iSTART, Point&Query, and AutoTutor. Project Listen [13] creates an environment where children and ESL (English as a second language) students can read text out loud with the aim to improve this skill. The software listens to the reader and makes suggestions on how to improve their reading skills. One of the ways that the software increases reading comprehension is by asking the students questions about the text that they just read [6]. Presumably, the increase in reading comprehension and word comprehension do not translate into helping the students enhance the deeper comprehension skills discussed by Anderson and Pearson [4] since this is not the aim of this particular software. The remaining environments, however, do take aim at creating deeper understanding within the reading comprehension field. Point&Query augments current learning environments, such as hypertext and hypermedia, by providing learner controlled question and answer sessions that expose readers to deep causal questions [10]. Both AutoTutor and iSTART make use of animated agents and natural language dialogue to scaffold inquiry strategies, metacognition, and explanation construction [10]. AutoTutor generates why, what-if, and how style questions and then enters into a dialogue with the student to expose the deeper constructs of the topic. iSTART takes a coaching approach to teach the students how to construct and improve self-explanations combined with other metacomprehension strategies. Although these systems have demonstrated student learning gains and improvement in learning strategies, more can still be done.

Many of the aforementioned tools, created to aid in reading comprehension, are more closed-ended systems that require a significant amount of time and energy to develop course content [15]. These closed-ended systems often make use of help requests to aid them in determining when a student is having a reading comprehension problem [6,10,13]. However, the vast

majority of environments such as WebCT/Blackboard, Moodle, etc., that are adopted by schools and post-secondary institutions are more open-ended in nature. These systems provide much more flexibility in terms of content development and improved ease in making changes to the content compared to what can be provided with closed ended systems. The problem with the open-ended systems is that they provide no real support for student learning other than providing the content for the students. However, open-ended systems do have good tracking facilities in place to capture student interaction with the system. By making use of current data mining techniques and pedagogy aimed at improving student learning, it is possible to capture students' cognitive behavior from these open environments.

Trace methodologies, such as capturing keystroke data, events, eye tracking data, etc., have demonstrated that data generated from a student's interaction with an environment can provide the necessary information to make cognitive and metacognitive interpretations [5]. This makes sense since how a student consumes content will have a direct effect on their comprehension of that content. If we know the following: what task the student is currently working on, the difficulty of the task, and the current behavior of the student as they work on the task, we can make cognitive interpretations [5,13,17,21]. Bloom's Taxonomy [3] of the Cognitive Domain, provides a pedagogical framework for determining how cognitively difficult a question/task is. Using this framework we can determine if the student's current cognitive skills are appropriate for the task that they are currently working on.

Bloom's Taxonomy [7] and its subsequent revision [3] are comprised of three overlapping domains: cognitive, affective and psychomotor. The affective domain is comprised of emotions, attitudes and values. The psychomotor domain is comprised of physical skill mastery, coordination, etc. The cognitive domain provides a method to classify educational objectives that relate to knowledge [21]. Within the cognitive domain are six hierarchical levels in order of increasing complexity. They are: knowledge, comprehension, application, analysis, synthesis and evaluation (as revised by Anderson et. al.[3]). The first three levels are considered to be foundational learning and are based upon the ability to know and apply factual knowledge [21]. The last three levels are considered higher level learning that is more abstract in nature [20]. Bloom had originally assumed that you could not achieve the higher levels without first mastering the lower levels of the hierarchy [7]. However, it appears that it is possible to work at the higher levels on some topics without first mastering the lower levels [3].

Wankat and Oreovicz [19] provide some examples of how to apply Bloom's taxonomy to an engineering domain. Knowledge or recall involves the descriptions, definitions, generalizations and other routine information about a topic. Comprehension involves understanding the technical representations of a topic including the translation, interpretation and extrapolation of that topic. Application involves the use of topical abstractions in explicit situations such as the use of rules, procedures and theories to perform some computation. Analysis involves breaking a problem into its principal parts in order to highlight any content hierarchy, properties. Furthermore, connections and structure found within the content are defined and clarified. Synthesis involves putting together all the constituent parts of a problem into a coherent system or solution. This can be very difficult since the process is open-ended and there may be many possible solutions to the problem. Lastly, evaluation can involve making conclusions about the value of materials used in a project or the methods used in that project. There is a need to satisfy specific criteria or use some standard of appraisal.

Through the use of the different levels of Bloom's Taxonomy and questions that are appropriately couched within the framework, it is possible to help learners to overcome the various problems originally posed by Anderson and Pearson [4].

2. METHODOLOGY

An experiment was designed to look for patterns of student behavior in a reading comprehension task. Students interacted with a learning environment designed to emulate hypermedia courses offered in post-secondary institutions where written content is presented along with questions about that content. The students could view the content and/or questions in any order or manner they chose with no constraints applied to their interaction with the system. In keeping with trace methodology approaches, all of the interactions/events with the content and questions were recorded and time-stamped. These would include events such as mouse click, mouse wheel, which item was clicked or selected and so on.

To aid in determining what part of the document was currently being read, a small scrollable text box that allowed 7 lines of text to be displayed as displayed in Figure 1. The size of the text box performed a couple of tasks. First, it does not take more than one minute to read the approximately 77 words contained within the text box. Although not directly used in the analysis, this could be used to determine if the individual was distracted from the task at hand. Second, it provided a means to determine how much time and how quickly the student read over the portions of the document that contained the answers to the various questions. The questions could be selected in any order and any text the students had entered into the answer text box was saved and displayed when the corresponding question was selected. None of the participants was observed, nor reported, as having any difficulty with operating the interface.

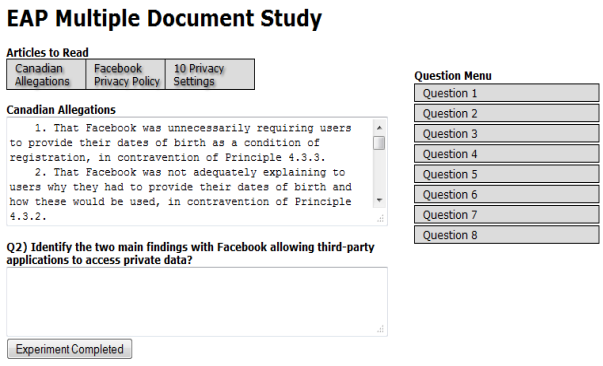

Figure 1 Screen Capture of Interface

The questions were developed using Bloom's Taxonomy Action Verbs [3,7]. Bloom and Anderson created a list of verbs that direct the way that a question should be answered. These verbs correspond to a level within Bloom's taxonomy. When you place the action verb at the beginning of the question, it frames the way that the question should be answered [7]. For example, Bloom's lowest level, knowledge, contains the action verb 'list'. Since the

task of the knowledge level is to remember previously learned information, successfully listing something that the student has previously read would demonstrate that the student has mastered that level of cognitive difficulty for that content. Questions at various Bloom levels were presented to the participants in a randomized order. All of the questions were present on the screen at all times and could be selected in any order by the student.

Questions were scored in order to provide a metric for how well the students were comprehending the content. In order to deal with the subjective nature of scoring question answers, a rubric was created according to the principles laid out in [16] for each of the questions. The rubric was revised a couple of times to take into account the various types of answers that were submitted during the testing phase of the development. For the lower Bloom levels, the answers generally came from one direct location within a document and so the scoring was fairly simple. For the higher level Bloom questions, information from multiple sources was expected. It was also expected that the students would bring their own prior knowledge to bear on the answer. It was here that revisions were required as the beta testing group interpreted the questions in unforeseen ways. The experiment was broken into two components. The first component provided the students with a single document to read and questions about that content were based upon the lower levels of Bloom's Taxonomy. All of the answers to the questions could be found within the document. The document that was chosen was a fairly technical document based upon Canadian privacy law as it applies to Facebook so that the participants would not have much prior knowledge of the specific subject matter. The participants were given 30 minutes to finish reading the document and answering four questions.

The second condition provided the students with two more documents in addition to the first document. The purpose of this condition was to better test the higher levels of Bloom's Taxonomy. The higher levels of Bloom's Taxonomy require synthesis and evaluation and so more information and documents were needed to allow for these requirements. The second document was instruction on how to implement advanced privacy features not commonly used within Facebook and the third document was a high level overview of the privacy settings used within Facebook. Again the answers to the high level questions could be found within the documents provided. However, in order to fully answer the higher level questions, information from more than one document was required. For this second condition, 90 minutes were allotted as the questions were more difficult and there were two new documents that needed to be read to generate complete answers. Two questions were aimed at the prior reading done in the first condition. One was a repeated question from the first condition and a second question was new but based solely on the information found in that first document. The remaining six questions were new and tested various levels of Bloom's Taxonomy. It was possible that the students could answer the questions in an increasing level of difficulty; however, they would have to purposefully select that order since the order in which they were presented was random.

Since the amount of time required to participate in both conditions might be a factor in participant involvement, both conditions were designed so that they could be run separately and using different participants depending on the participants' wishes. In the actual running of the experiment, the majority of the participants moved from the first condition right into the second condition with no delay. The participants were adult students enrolled in a grade 12 Saskatchewan Institute of Applied Science and Technology (SIAST) Adult Education English course. There were 17 participants for the first condition and 11 for the second condition with an average age of 26.

3. RESULTS

The 28 participants generated over 8500 events in total from both conditions. Events such as the mouse clicking on a specific button or object and mouse wheel scrolling were captured. Each event was time-stamped with the user-id, event-id, current question-id, current document-id, and position within the current document. This gave us what task/question the student was currently working on, which document they were working on, where in the document they were, and what event they were using. For example, if the student moved the scroll wheel of the mouse to move down in the document we could then determine from the time-stamp data and the position data, how quickly and what material they were reading. With this information we can begin to deduce student behavior as they work at completing the various questions.

In order to determine how much reading the students were doing, the timestamp data was processed so that reading, scanning and scrolling navigation times could be calculated for each interaction/event. The time cutoffs used to distinguish reading from scanning from scrolling fit with other document navigation research [1]. Any time between events greater than five seconds was classified as reading. Any time greater than two seconds but less than five seconds was classified as scanning and any time less than two seconds was classified as scrolling. The reading time also encompassed time that the participant spent thinking about the answer. In the 8500 events captured across the 28 participants, only 13 events had a time greater than two minutes and only 33 events had a time greater than one minute before another event was performed. Given the time it takes to read the content in the textbox, the total time between events including the reading and thinking times, was not a large enough percentage of the data to warrant separate classification.

The total amount of time that a student spent in the experiment was calculated and used to create a ratio of time spent by the student reading, scanning and scrolling. This ratio was then broken down into the reading, scanning and scrolling ratios for each individual question. When combined with the level of difficulty for each question, as determined by Bloom's taxonomy, it was possible to tie student reading behavior to the difficulty of the task.

In order to see if there were students who behaved similarly for different levels of difficulty, we implemented the Forgy method for k-means clustering for d=3 dimensions and k=4 [11]. Hammerly et al. [11] demonstrated that the Forgy method for initialization was the preferred method for initializing the standard k-means, also known as Llyod's, algorithm [12]. The dimensions that we chose were the reading, scanning, and scrolling axes. K = 4 was chosen since our sample size was small. More than 4 clusters produced some clusters where there were too few to be statistically analyzed. Since the algorithm randomly chooses its centroid points, there is no researcher bias entering into the initial sets of clusters that were created. In order to find as many interesting clusters as we could, the Forgy k-means algorithm was iterated multiple times. We defined interesting clusters as those clusters that elicited either positive or negative reading, scanning or scrolling behaviors. A positive behavior is defined as a

behavior that results in a good grade. A negative behavior is defined as a behavior that results in a poor grade. Those clusters that presented with both positive and negative behaviors were deemed less interesting. Each time an interesting cluster was found, the centroid was recorded. Once multiple interesting centroids were found, the most interesting centroid found was hard coded as a starting centroid. The hard coding of the algorithm removes one of the random initializations from the Forgy initialization and inserts the most interesting centroid in its stead. For example, the experiment used k = 4 random clusters in the initialization. With the hard coded cluster added, k = 3 random and k = 1 hard coded are what the algorithm would initialize with. The algorithm was run again with one hard coded centroid and three randomly chosen centroids to see how the other random clusters interacted since how the cluster is initialized is known to have an effect on how the other clusters form [11]. If a new interesting cluster was discovered that was more predictive of students' behavior than a previous closely related centroid, the old centroid was removed in favor of the new centroid. If no more centroids were discovered that were more interesting than the hard coded centroid, then the second most interesting centroid was hard coded and the remaining two centroids were left random and the above process was duplicated with two hard coded centroids. A third hard coded cluster was added in accordance with the above procedure and the process was performed again until all four of the initializations centroids were hard coded.

Over multiple iterations six interesting clusters were discovered with two of these clusters containing too few data points to be included in any statistical analysis that was performed. The following clusters proved to be statistically interesting with respect to the Bloom level:

- Light Reading Cluster: 50% reading: 30% scanning: 20% scrolling (50:30:20)
- Light Medium Reading Cluster: (60:30:10)
- Heavy Medium Reading Cluster: (70:20:10)
- Heavy Reading Cluster: (80:10:10)

The two clusters, Medium Scrolling (10:10:60) and Medium Scanning (10:60:10), were clusters that we expect to play a more important role in future experiments. However, due to our sample size, they could not be used in our statistical analysis.

An ANOVA was performed on each of the clusters to see if a statistically significant relationship could be found between the different reading behaviors as clustered by k-means and the Bloom levels of the questions within the experiment. The tests were performed at the $\alpha = 0.05$ level. Questions at Bloom levels 1,2,3,5, and 6 were provided in this experiment. There were no Bloom level 4 questions, to give learners time to answer more questions at level 5 and 6 within the overall time constraints.

Table 1 shows that, with the exception of level 5, all of the Bloom levels were statistically significant. The null hypothesis used for these tests are that the means for each of the clusters does not vary according to the Bloom level that is being tested. In other words, the reading, scanning and scrolling means should be the same for all of the clusters found by k-means. Table 1 shows that the differences found between the clusters for each of the Bloom levels were not due to random chance. The p-values indicate that, in all but two cases, there is a really small chance of getting these results if no real difference between the groups exists. This indicates that the students' reading, scanning and scrolling behaviors captured by the system and then clustered are significantly different from one another as it relates to the level of Bloom's taxonomy. For example, those students who were classified as Light Readers based on the reading, scanning and scrolling ratios for Bloom level 1 were significantly different from those who were classified as Light Medium Readers for the same Bloom level. However, the ANOVA itself cannot make this exact determination of which cluster is significantly different from another cluster; it can only tell us that there is a significant difference between some of the groups in the analysis. Further analysis, discussed later on, is needed in order to see which of the clusters are significantly different from each other.

Although inclusion in a cluster does not completely predict scores, it is indicative of overall performance. For example, take question 2 in the first condition (low level Bloom with a single document) that was designed to force the students to scan through the document as they needed to count the number of instances that a certain event, such as a successful appeal on a complaint about Facebook to the Canadian Privacy Commission, occurred in the document. This type of problem is often present in many forms in academia and the work place where it is necessary to arrive at a solution within the time constraints. 100% of the students in the Light Reading (50:30:20) cluster, which was higher in scanning and scrolling times, achieved full marks or close to full marks. Correspondingly, those students in the Heavy Reading (80:10:10) cluster scored no better than 50% with over ½ of the students in the cluster scoring 0%. Since the source materials were present for the duration of the experiment and there were time constraints, the Heavy Reading strategy is not the best strategy to be used in this situation. This result is somewhat surprising since it is generally accepted that Heavy Reading is considered a good cognitive strategy in a reading comprehension task. In this case, the cognitive load required to be able to answer this type of question, the time limitations of the experiment and the fact that the source materials were available, make the adoption of the Light Reading strategy a better choice. The reduction in the cognitive load by choosing to perform more scanning and scrolling through the document rather than committing the information to memory when performing Heavy Reading allows the participants to perform better on this type of task. It should be noted that for other tasks, a Heavy Reading strategy is the best choice. Furthermore, in situations where the source materials are not available during the task, the Heavy Reading strategy is most likely the best choice regardless of the task given.

The Heavy Reading strategy proved to be the most successful strategy as the level of difficulty for the questions increased as measured by Bloom's Taxonomy. The participants were able to achieve better marks compared to those that chose a Light Medium Reading strategy. For example, question six of the second condition required the participants to put together various thoughts and ideas about FaceBook privacy policy from multiple documents into a complete whole thought that did not exist in any of documents (Bloom level 6). For this problem the students fell into multiple clusters. Each document had its own set of events that tied the reading, scanning and scrolling ratios to that document. This provides a mapping of how each student used each document to answer a particular question. In order for the students to get a good grade they needed to fall into the Heavy Reading category on all the documents that were required to fully answer the question. Those students that performed Heavy Reading on all the necessary documents scored well. The students that performed Heavy Reading on only one of documents

they were required to read did not score above 30%. Those students that performed Heavy Reading on two of the required documents scored no higher than 83% and those that performed Heavy Reading on all of the documents scored no lower than 83% and up to 100%. Those students that used the Light Medium Reading strategy scored 0%. There was one student who scored 30% that used the Light Medium Reading strategy but their answer contained no content from any of the documents, rather they used extraneous information from their previous experience.

The use of the Light Reading strategy did not appear above Bloom level three and the Light Medium Reading strategy appeared throughout the Bloom levels. At Bloom level's five and six, those participants that chose to use the Light Medium Reading strategy did not receive good grades. The availability of the source documents to the participants did not aid them in answering more challenging questions. The participants needed to be able to recall information from a variety of sources in order to be able to fully answer the questions. Instead of using source material, probably because they could not recall where it was or if it was present, they used incorrect information from some other source outside of the experiment. It should be noted that they did not access supplementary material from either books or the Internet during this experiment.

These aforementioned patterns of behavioral clustering being predictive of marks do not hold in all cases. For example, the Light Reading (50:30:20) cluster for question 3 in the second condition (higher level Bloom with multiple documents) had 50% of the students achieve 100% while the other 50% received 0%. Since we captured the current reading position within the document with each event, we can determine the amount of time spent reading, scanning and scrolling over the portions of the document that contain the answer. When analyzed, this information is able to fully account for the differences in scores found within the cluster of the above example. For example, those students who received 0% spent the majority of their time scrolling and scanning compared to those who received full scores, who spent much more time reading over the portion of the document that contained the answer.

The Low Level and High Level analysis from Table 1 shows that when Bloom is broken down into two categories, low level (levels 1, 2, and 3) and high level (levels 5 and 6) there are significant differences for both the high and low levels between the clusters. When we perform the Tukey-Kramer test later on (Table 5), it will show that all of the clusters are significantly different from one another as well. Interestingly, when we combine all the levels together to see if the clusters by themselves are statistically different, we get no significant results. In other words, higher level and lower level meta-cognitive reading strategies seem to elicit different behavior on the part of learners.

In order to find out which clusters were significant from each other, a Tukey-Kramer analysis is required. A Tukey-Kramer analysis allows a pairwise comparison of each of the clusters and allows a comparison of groups that do not have the same number of students. The minimum significant difference value was used to calculate if the pairwise comparison was significant and correct for the multiple comparisons. The numbers in the top right hand portion of the Tables 2 through 5 show the Tukey-Kramer minimum significant differences (MSD). The numbers in the lower left corner of Tables 2 through 5 show the observed absolute value of the difference in means between each pair of groups. Those numbers in the lower left of the tables marked with an asterisk are deemed significant if they are larger than their corresponding MSD located in the top right of the table. Table 2 shows that all of the clusters were significantly different from each other. This was found for all of the other Bloom levels except for Bloom level 2 and 5. Table 3 shows that there are significant differences between most of the groups except for the Medium Heavy Reading cluster and the Heavy Reading cluster for Bloom level 2. Although the k-means algorithm clustered these reading, scrolling and scanning ratios into two different clusters, the actual differences between the ratios was close. So the grades tended to be higher in the Medium Heavy Reading and at the same time lower in the Heavy Reading cluster. It was situations like this one where the ratios were close together that made us wonder if a breakdown of individual Bloom levels was the best predictor or if the levels should be more coarse-grained and moved into a high level Bloom category and a low level Bloom category rather than individual Bloom levels.

One of the major problems with this experiment was that we did not have a large enough sample size for the higher levels of Bloom as tested in the second condition. Table 4 shows that there were no significant differences found between any of the clusters at Bloom level 5. A more in-depth analysis showed that most of the students chose a similar strategy to answer those questions and

Bloom Level	F	P	F-Critical
1	79.94	3.14E-16	2.86
2	39.31	3.74E-11	2.88
3	147.93	4.80E-11	3.63
5	0.60	0.63	3.59
6	50.77	0.000385	5.99
Low Level	209.48	1.83E-43	2.48
High Level	95.95	1.64E-17	2.86
All Levels Combined	1.40	0.25	2.68

Table 1 One way ANOVA for Bloom Level

	50,30,20	**60,30,10**	**70,20,10**	**80,10,10**
50,30,20	-	0.08311	0.07745	0.08089
60,30,10	0.16204*	-	0.07976	0.08311
70,20,10	0.2963*	0.13426*	-	0.07745
80,10,10	0.4447*	0.28268*	0.14842*	-

Table 2 Tukey-Kramer Analysis Bloom Level 1 (* denotes significant differences)

	50,30,20	**60,30,10**	**70,20,10**	**80,10,10**
50,30,20	-	0.21238	0.19337	0.17629
60,30,10	0.21341*	-	0.2324	0.21839
70,20,10	0.4906*	0.2772*	-	0.19995
80,10,10	0.6724*	0.459*	0.18183	-

Table 3 Tukey-Kramer Analysis Bloom Level 2

	60,30,10	70,20,10	80,10,10
60,30,10	-	40.5993	28.4136
70,20,10	12.182	-	38.9069
80,10,10	11.666	0.5159	-

Table 4 Tukey-Kramer Analysis Bloom Level 5

	50,30,20	60,30,10	70,20,10	80,10,10
50,30,20	-	0.09225	0.07145	0.0668
60,30,10	0.15755*	-	0.09473	0.09127
70,20,10	0.27206*	0.11451*	-	0.07019
80,10,10	0.4222*	0.26463*	0.15012*	-

Table 5 Tukey-Kramer Analysis Low Level Bloom

therefore no significant differences were found between the clusters. With a larger sample size we believe that even this would become statistically significant. The fact that a significant difference was found in Bloom level 6 may just be due to an artifact in the data; however, the significant differences found in the lower Bloom levels 1 through 3, given the slightly larger N, seems to imply that with a larger N we will see those same differences in the higher Bloom levels.

For the low level and high level Bloom groupings, significant differences were found between all of the clusters. Table 5 shows the significant differences found for the low level Bloom grouping.

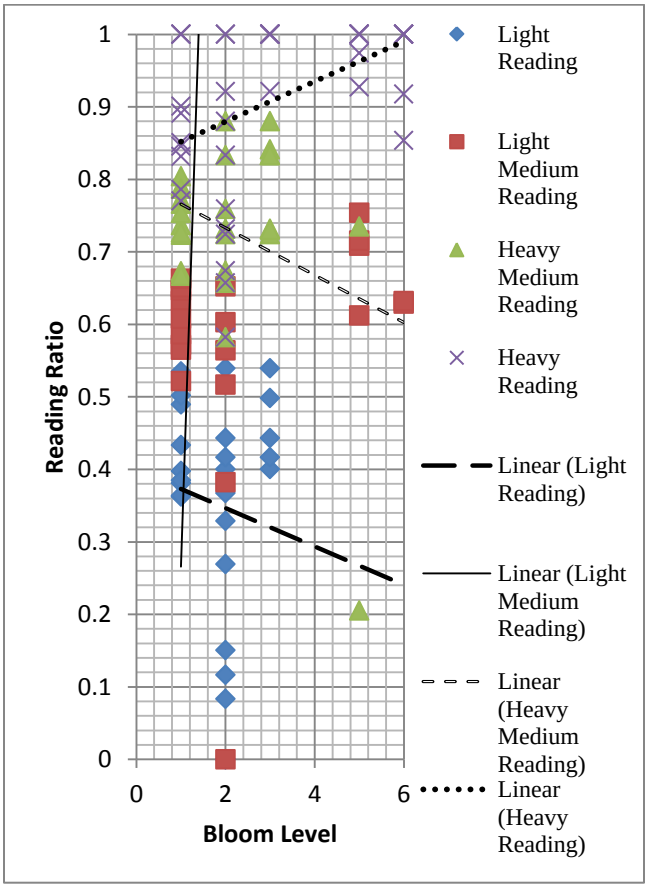

Figure 2 Graph of Reading Ratio vs Bloom Level

Next we analyzed how the clusters were related to the Bloom levels. Figure 2 shows that the Light Reading behavior was not found in any questions above Bloom level 3. This seems to indicate that Light Reading behavior is not conducive to the more cognitively difficult tasks. The Heavy Medium Reading cluster had only 2 instances in questions above Bloom level 3. The decreasing use of Heavy Medium Reading as the Bloom level of difficulty increases shows that some of the students' adapted a heavier reading behavior compared to the Light Reading behavior at the lower Bloom levels. They gave up the Heavy Medium Reading strategy for the Heavy Reading strategy used more in the higher Bloom levels. The Heavy Reading cluster was found at each of the Bloom levels. As the Bloom levels increase in difficulty, the amount of Heavy Reading increases until all but one student are Heavy reading at Bloom level 6. Correspondingly, the Light Reading cluster that contains more scrolling and scanning decreased as the Bloom level increased. This seems to suggest that different strategies are appropriate for different Bloom levels. There were several students that used the same cognitive strategy throughout the experiment despite the difficulty of the tasks changing. For example, some of the students chose a Heavy Reading strategy for the entire experiment. As a result they did not complete the experiment with respect to answering all of the questions as they spent too much time reading and not enough time answering the questions. Furthermore, students who chose a Heavy Reading strategy for the lower level Bloom questions did not always score very well even though the questions were cognitively simpler. The question 2 example from the first condition cited earlier in the paper is a good example. Other students chose different strategies for different levels of difficulty. For example, they would choose a strategy that was higher in scanning for the lower levels of Bloom and switch to a Heavy Reading strategy at the higher levels of Bloom. These students were able to complete the experiment and answer all of the questions within the time allotted.

At Bloom level 6 only two strategies are used: the (60:30:10) Light Medium Reading and the (80:10:10) Heavy Reading strategies. Although the students' inclusion in the Heavy Reading cluster was a good indicator of higher scores, there was still a lot of variance in the grades found within the Heavy Reading cluster for Bloom level 6. The best predictor of scores within the cluster was the ratio of reading time spent over the position in the various documents that contained the material necessary for the answers. This helped to identify those students that merely used their own unsupported opinions to answer questions versus those students that used information from the articles to support their answer.

4. CONCLUSIONS

This experiment demonstrates that the various cognitive strategies used by students to solve tasks of varying degrees of difficulty can be recognized automatically by an ITS. The use of Bloom's Taxonomy for categorizing the difficulty of the task and k-means clustering on the reading-scanning-scrolling strategies allowed for the detection of these cognitive strategies. These clusters can easily be turned into metrics that can be used by a system to discover the strategies the students are using and provide the necessary metacognitive suggestions to improve the student's cognitive skill set. Furthermore, the experiment shows that students may not always select the best strategy to use. This approach is not refined enough to predict an actual score on a question. However, it does provide a method of determining the reading strategy being used and predicting if the cognitive

strategy that is being employed is one that is positive or negative given the difficulty level (in terms of Bloom) of the question. Furthermore, since we are able to detect these inconsistencies in the use of cognitive strategies automatically we have the potential to automatically update a student model, and thereby inform the student about the metacognitive strategies they are employing and/or suggest appropriate pedagogical tasks that could be useful for a student attempting to improve weak metacognitive skills in the reading comprehension domain at least.

It is possible that the course grained detection of cognitive strategies will provide direction for systems where the application of more fine grained searches and algorithms might be able to predict the grade or allow for some specific pedagogical interventions. For example, do the students perform some course grained strategy in their initial search through a document and then use that information to refine their strategy for one that is more optimal for the solving some particular task?

K-means clustering comes with its benefits and drawbacks. The benefit of this algorithm is that it arrived at four interesting centroids that are hard-coded and that can be used in a real-time algorithm for the detection of significant reading strategies. There are other clustering methods, such as EM clustering, that may work better at determining new cluster centroids or are better at including the students in the correct cluster. This will be a subject of further research.

Future experiments also need to be performed to increase the sample size of the experiment, especially in terms of the higher Bloom levels. The increased sample size should allow us to see statistically significant cognitive skill differentiation at the higher Bloom levels but should also help to validate the reading-scanning-scrolling clusters that were not statistically viable with the current sample size. These experiments should further help solidify the use of Bloom's Taxonomy as a tool in detecting cognitive strategies for reading comprehension tasks. Furthermore, the interplay between reading comprehension and document selection may provide some interesting insights at the higher levels of Bloom's Taxonomy.

5. ACKNOWLEDGMENTS

Our thanks to Canada's Natural Sciences and Engineering Research Council for helping to fund this research and to those SIAST students who volunteered to take part in the experiment.

6. REFERENCES

[1] Alexander, J., and Cockburn, A. (2008). An empirical characterization of electronic document navigation. *Proceedings of Graphics Interface '08.* Windsor, On., Canada, 123-130.

[2] Alfassi, M. (2004). Reading to learn: Effects of combined strategy instruction on high school students. *Journal of Educational Research*, 97, 171–184.

[3] Anderson, L.W., Krathwohl, D.R. and Bloom, B.S., (Eds), 2000, Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman.

[4] Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson(Ed.), Handbook of reading research (pp. 829-864). New York: Longman.

[5] Azevedo, R., Moos, D., C., Johnson, A., M., Chauncey, A., D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist*, 45(4), 210-233.

[6] Beck, J. E., Mostow, J., Cuneo, A., & Bey, J. (2003). Can automated questioning help children's reading comprehension? *Proceedings of the Tenth International Conference on Artificial Intelligence in Education* (AIED2003), Sydney, Australia

[7] Bloom, B.J., Englehart, M.D., Furst, M.D., Hill, E.J. and Krathwohl, D.R., (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain.* New York: David McKay.

[8] Bransford, J. D., Stein, B. S., Nye, N. J., Franks, J. F., Auble, P. M., Merynski, K. J., & Perfetto, G. A. (1982). Differences in approaches to learning: An overview. *Journal of Experimental Psychology: General*, 3, 390-398.

[9] Fischer, C. (2003). Revisiting the reader's rudder: A comprehension strategy. *Journal of Adolescent & Adult Literacy*, 47, 248–256.

[10] Graesser, A. C., McNamara, D. S. & VanLehn K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist*, 40(4), 225-234.

[11] Hammerly, G., and Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. *Proceedings of the ACM Conference on Information and Knowledge Management*, 600-607.

[12] Lloyd, S. P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd., S. P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory 28 (2): 129–137.

[13] Mostow, J., & Chen, W. (2009). *Generating Instruction Automatically for the Reading Strategy of Self-Questioning.* Proceedings of the 14th International Conference on Artificial Intelligence in Education, 465-472.

[14] Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, 36, 89–101.

[15] Razzaq, L., Patvarczki, J., Almeida, S.F., Vartak, M., Mingyu Feng, Heffernan, N.T., Koedinger, K.R. (2009) The ASSISTment Builder: Supporting the life cycle of tutoring system content creation, *Learning Technologies, IEEE Transactions on* , vol.2, no.2, pp.157-166.

[16] Schraw, G., (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist*, 45(4), 258-266.

[17] Thompson, E., Luxton-Reilly, A., Whalley, J. L., Hu, M., and Robbins, P. (2008). Bloom's taxonomy for CS assessment. *Proceedings of the Tenth Conference Australasian Computing Education*, 155-161.

[18] van Keer, H. (2004). Fostering reading comprehension in fifth grade by explicit instruction in reading strategies and peer tutoring. *British Journal of Educational Psychology*, 74, 37–70.

[19] Wankat, P., and Oreovicz, F., (1993). *Teaching Engineering.* New York: McGraw-Hill.

[20] Yao-Ting, S., Kuo-En, C., & Jung-Sheng, H. (2008). Improving children's reading comprehension and use of strategies through computer-based strategy training. *Computers in Human Behavior*, 24, 1552-1571.

[21] Zywno, M., S. (2003). Hypermedia instruction and learning outcomes at different levels of Bloom's Taxonomy of Cognitive Domain. *Global Journal of Engineering Education*, 7(1), 59-70.