Doctoral Consortium

Towards the Understanding of Gestures and Vocalization Coordination in Teaching Context

Roghayeh Barmaki
Department of Computer Science
University of Central Florida
barmaki@cs.ucf.edu

Charles E. Hughes
Department of Computer Science
University of Central Florida
ceh@cs.ucf.edu

ABSTRACT

Nonverbal behaviors such as facial expressions, eye contact, gestures, postures and their coordination with voice tone and prosody have strong impact on the process of communicative interactions. Successful employment of nonverbal behaviors plays an important role in interpersonal communication in the classroom between students and the teacher. Student teachers need to improve their teaching skills, from communication to management, and prior to entering the classroom. To support these aspects of teacher preparation, we developed a virtual classroom environment, TeachLivE^{TM} for teacher training, reflection and assessment purposes. In this work we investigate the connections between gestures and vocalization characteristics of participants in a teaching context for two settings within the TeachLivE environment.

We have developed an immediate feedback application that is presented to the participants in one of the study settings. It provides visual cues to the participant in front of the tracking sensor any time that she exhibits a closed stance. Identification of these type of connections between acoustic and gestural components of communication provides an added dimension that could assist us in using machine learning methodologies to extract multimodal features as teaching competency measures.

Keywords

gesture; vocalization; nonverbal behavior; Microsoft Kinect; virtual teaching rehearsal environment.

1. INTRODUCTION

Interpersonal communication involves a variety of modes and components in communication. We might think that actual words are the primary part of communication; however, the majority of interaction between individuals, including students and teachers, is nonverbal, encompassing between 65 and 93 percent of what occurs related to learning [7]. These nonverbal elements include both nonvocal (e.g. body language) and vocal components (e.g. voice pitch and intonation). Body language by itself include several aspects: facial expressions, eye contact, posture or stance, gestures, touch and appearance. This research investigates the connection of postures and/or gestures with acoustic components of the nonverbal communication in the teaching context.

Multimodal analysis co-processes two or more parallel input streams (modes) from human-centered interactions that contain rich high-level semantic information [9]. Teaching and learning have always been multimodal as both are unified with speech, gesture, writing, image and spatial setting [12]. Multimodal data analysis in a teaching context helps us to have an informed understanding of the performances of the teacher participants.

TeachLivE is a simulated classroom setting used to prepare teachers for the challenges of working in K-12 classrooms. Its primary use is to provide teachers the opportunity to rehearse their classroom management, pedagogical and content delivery skills in an environment that neither harms real children, nor causes the teacher to be seen as weak or insecure by an actual classroom full of students. TeachLivE uses its underlying multi-client-server architecture called AMITIES- Avatar Mediated Interactive Training and Individualized Experience System [8]. A human-in-the loop (called an interactor) orchestrates the behavior of the virtual students in real-time based on each character's personality and backstory, a teaching plan, various genres of behaviors and the participant's input. The virtual classroom is displayed on a large TV screen to the participant and the view of the virtual classroom scene dynamically changes based on the participant's movements in front of the tracking sensor. We have developed a real-time gesture recognition application for nonverbal communication skill training, based on the Microsoft Kinect SDK [1] as part of ReflectLivE, the TeachLivE integrated reflection tool [3]. The hypothesis is that our developed feedback application has positive impact on the participants' body language, leading to more open and fewer closed stances. The open stance has arms and legs not crossed in any way. To explore the validity of this hypothesis and system usability evaluation, we report the results from the conducted case study with two settings using the feedback application (section 2.1).

We are also interested in looking at the connections between the participant's gestures and acoustic characteristics in different situations in the classroom, such as while asking questions from virtual students, conversation turn-taking after students' responses, introducing a new topic, etc. The analysis of the recorded sessions from a gesture-voice aspect is another motivation for this research that seeks a broader understanding of communication practices that reflect and support teaching competency.

Investigating the related research, there have been a number of prior attempts to develop social skill training and

feedback applications using interactive environments. Presentation Trainer [10] collects multimodal data using the Microsoft Kinect and provides immediate cues about the trainee's body posture, embodiment and voice volume during her presentation. Similarly, Dermody and Sutherland [5] present a multimodal prototype for public speaking purposes that uses the Kinect sensor. Their system provides real-time feedback on gaze direction, body pose and gesture, vocal tonality, vocal dysfluencies and speaking rate.

At first glance, gesture and speech may be coupled less directly than, e.g., prosody and speech, as both originate in very different physiological systems. However, some views and findings suggest a close connection between both, especially in production. This mutual co-occurence of speech and gesture reflects a deep association between the two modes that transcends the intentions of the speaker to communicate [11].

2. APPROACH

We present our research to understand the gesture and vocalization connections in the following two separate subsections since most of our currently reported research has been done independently with our effort to fuse the collected multimodal data still under development.

2.1 Gesture

This research evolved based on the existing literature expressing the importance of open body gesturing in successful interactive teaching (teaching competency) [2]. Reviewing the existing recordings of teaching sessions in TeachLivE gave us a baseline about the way teachers use their body in the virtual classroom. In our observations, most of the teachers were not thoughtful of their body movements and many of them exhibited closed stances most of the time in their teaching sessions. The recognized frequent closed postures (or closed gestures) were hands folded in front and back, hands on hips, and crossed arms. These gestures are noted as closed or "not-recommended" gestures. We are interested in detecting these closed gestures and reminding the trainees about their closed body language. In social skill training, the impact of immediate and real-time feedback in the rehearsal process has been reported as very positive in comparison to other types of feedback provision such as delayed feedback [10]. The developed feedback application is capable of providing visual or haptic (vibration wrist band) prompts in real-time for targeted closed gestures. The effectiveness of the implemented visual feedback application was evaluated by conducting a user study. It was a single-time within-subjects, counterbalanced study with two settings (TeachLivE with and without feedback application) and each session was 7-minute long. Participants (N=30, 6M, 24F) were asked to attend both of the settings, and complete pre and post questionnaires (the total recruitment time was approximately 45 minutes per participant). We randomly assigned the participants into two groups A and B, where group A (N=15, 3M, 12 F) experienced TeachLivE with feedback setting in their second session and group B had this experience in their first session. The collected full-body tracking data from the participants was processed [3] to extract the percentage of time that a subject exhibited closed gestures (CGP) in the recorded sessions. Our expectation based on the hypothesis (section 1) was that we would observe a considerable difference between groups A and B in the first session and a slight difference between the two groups in the closed body gesture employment in the second session. To evaluate the impact of our proposed feedback application on body language thoughtfulness, we calculated CGP for 60 recorded clips from 30 participants. The boxplot in Figure 1 presents the distribution of CGP between two groups of participants.

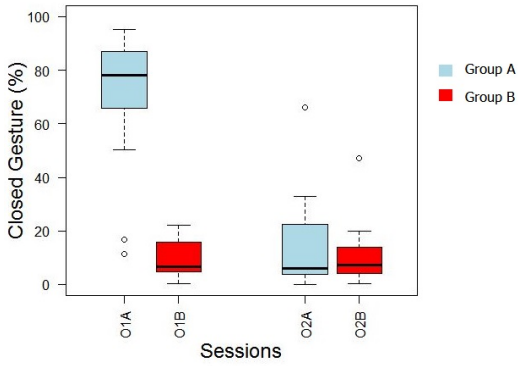

Figure 1: Medians and interquartile ranges of CGP exhibition in two sessions (observations) among groups A and B. Circle represent outliers.

Figure 1 shows some of key findings from this study. It presents the wide range (from 95% to 16%) of closed gesture employment for group A in the first session. It also indicates the median of CGP for group B participants is lower than group A (6.4 % and 7.2% for two sessions for group B and 78% and 5.9% for group A). As Figure 1 indicates, the hypothesized statement is supported for the participants of the study. The average time that all of the participants in group A exhibited closed gestures reduced significantly from their first session to their second session. Most interestingly, the participants in group B exhibited open gestures most of the time even in the second unaided session.

2.2 Vocalization

In this study, we recorded video, audio, full body tracking data and event logging information (including virtual students' talk-time and behaviors) from the TeachLivE system. The reader can find further recording details in [3].

After collecting the data, we processed the recorded audio from video sessions using Audacity software to extract the Waveform Audio File Format from recorded avi files. We opened the .wav files in the Praat tool [4] and extracted some basic vocal characteristics (pitch and intensity objects) from the audio files. Praat is a free computer software package for the analysis of speech. Voice pitch is the perceptual correlate of vocal fundamental frequency and voice intensity indicates voice loudness in db. A PitchTier object represents a time-stamped pitch contour (hereby feature), i.e. it contains a number of (time, pitch (Hz)) points, without voiced/unvoiced information. An IntensityTier object represents a time-stamped intensity contour, i.e., it contains a series of (time, intensity) points [4]. Pitch and intensity tier associated with our recorded sessions were exported for multimodal analysis purpose to the ANVIL [6]. ANVIL is a video annotation tool that offers multi-layered annotation

based on a user-defined coding scheme. Figure 2 shows the ANVIL tool.

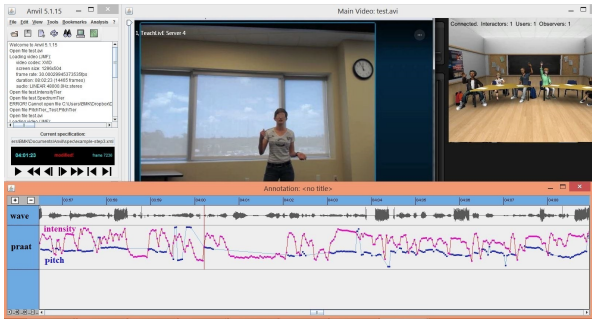

Figure 2: TeachLivE video sessions (including the participant front view and virtual classroom scene) within the ANVIL annotation tool [6]. Three acoustic contours waveform, pitch (blue) and intensity (pink) [4] are imported to the annotation project.

We intend to add our gesture recognition application output as an extended contour in the ANVIL. This will automatically present the types and timing for different closed gestures during the recorded session. The current version of the ANVIL does not support the exported (closed) labels of frames from the Kinect V2 gesture recognition tool as a contour, so we are working on this open-source tool to develop our desired contour structure. As mentioned earlier, our goal of using ANVIL is to understand the correlations of acoustic features with gesturing in these three main cases: 1) when the participant teacher asks a question from virtual classroom, 2) when the teacher listens to the responses from the class (conversation turn taking between students and teacher), and finally 3) when the teacher introduces a new or abstract topic or is summarizing the discussion. Literature supports that teachers gesture more in the mentioned cases [2]. We will annotate the recorded videos based on the teaching plan, conversational cases, open/closed, and affirmative gesture employment. The automatically generated vocalization information would be exported in conjunction with manual annotation data for further analysis.

3. CLOSING REMARKS

The study reported here fills a gap in multimodal research for education. In this paper, we first explained the impact of nonverbal behaviors in teaching competency. We then reported a case study to evaluate the performance of our developed feedback application for nonverbal communication skill training. We used the Microsoft Kinect sensor and its full-body tracking data stream to develop our real-time gesture feedback application. The results from the recorded body tracking data indicated the positive impact of informed body language and gesture in communication proficiency. We also introduced relevant tools and techniques for multimodal feature extraction for teaching competency, and we expect to report the results after developing an appropriate coding scheme framework and the annotation procedure.

For future research, we are looking forward to uncovering additional teaching evaluation insights with the analysis and evaluation of multimodal recorded data, as multimodality is an integral part of teaching.

Acknowledgments

The authors acknowledge the support of the Bill & Melinda Gates Foundation (OPP1053202) and the National Science Foundation (CNS1051067, IIS1116615). We also wish to express our gratitude to the entire TeachLivE team, especially the interactors who give life and authenticity to our avatars. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

4. REFERENCES

[1] Visual gesture builder: A data-driven solution to gesture detection. http://aka.ms/k4wv2vgb, July 2014. Retrieved 3/10/2016.

[2] M. W. Alibali and M. J. Nathan. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. *Video research in the learning sciences*, pages 349–365, 2007.

[3] R. Barmaki and C. E. Hughes. Providing real-time feedback for student teachers in a virtual rehearsal environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 531–537, New York, NY, USA, 2015. ACM.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program] version. 6.0.17, 2016. Accessed 5/07/2016 from http://www.praat.org/.

[5] F. Dermody and A. Sutherland. A multimodal system for public speaking with real time feedback. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 369–370, New York, NY, USA, 2015. ACM.

[6] M. Kipp. Anvil: The video annotation research tool. In *The Oxford Handbook of Corpus Phonology*. Oxford University Press, 2014.

[7] A. Mehrabian. *Silent Messages: Implicit Communication of Emotions and Attitudes*. Wadsworth, 1972.

[8] A. Nagendran, R. Pillat, A. Kavanaugh, G. Welch, and C. Hughes. A unified framework for individualized avatar-based interactions. *Presence: Teleoper. Virtual Environ.*, 23(2):109–132, Aug. 2014.

[9] S. Oviatt and P. R. Cohen. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*. Morgan & Claypool Publishers, 2015.

[10] J. Schneider, D. Börner, P. van Rosmalen, and M. Specht. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 539–546, New York, NY, USA, 2015. ACM.

[11] P. Wagner, Z. Malisz, and S. Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014.

[12] M. Worsley, K. Chiluiza, J. F. Grafsgaard, and X. Ochoa. 2015 multimodal learning and analytics grand challenge. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 525–529, New York, NY, USA, 2015. ACM.

Towards Modeling Chunks in a Knowledge Tracing Framework for Students' Deep Learning

Yun Huang
Intelligent Systems Program
University of Pittsburgh
210 S. Bouquet Street
Pittsburgh, PA, USA
yuh43@pitt.edu

Peter Brusilovsky
School of Information Sciences
University of Pittsburgh
135 N. Bellefield Ave.
Pittsburgh, PA, USA
peterb@pitt.edu

ABSTRACT

Traditional Knowledge Tracing, which traces students' knowledge of each decomposed individual skill, has been a popular student model for adaptive tutoring. Unfortunately, such a model fails to model complex skill practices where simple decompositions cannot capture potential additional skills that underlie the context as a whole constituting an interconnected *chunk*. In this work, we propose a data-driven approach to extract and model potential *chunk units* in a Knowledge Tracing framework for tracing deeper knowledge, which is primarily based on Bayesian network techniques. We argue that traditional prediction metrics are unable to provide a "deep" evaluation for such student models, and propose novel data-driven evaluations combined with classroom studies in order to examine our proposed student model's real-world impact on students' learning.

Keywords

complex skill, chunk, robust learning, deep learning, Knowledge Tracing, Bayesian network, regression

1. INTRODUCTION

Knowledge Tracing (KT) [4] has established itself as an efficient approach to model student skill acquisition in intelligent tutoring systems. The essence of this approach is to decompose domain knowledge into elementary skills, map each step's performance into the knowledge level of each single skill and maintain a dynamic knowledge estimation for each skill. However, KT assumes skill independence in problems that involve multiple skills, and it is not always clear how to decompose the overall domain knowledge. Recent research demonstrated that the knowledge about a set of skills can be greater than the "sum" of the knowledge of individual skills [8], some skills must be integrated (or connected) with other skills to produce behavior [11]. For example, students were found to be significantly worse at translating two-step algebra story problems into expressions (e.g., 800-40x) than they were at translating two closely matched one-step problems (with answers 800-y and 40x) [8]. Also, recent research that has applied a difficulty factor assessment [1] demonstrated that some factors underlying the context combined with original skills can cause extra difficulty, and should be included in the skill model representation. Meanwhile, research on computer science education has long argued that knowledge of a programming language cannot be reduced to simply the "sum" of knowledge about different constructs, since there are many stable patterns (schemas, or plans) that have to be taught or practiced [16]. We summarize the above findings and connect them with a long-established concept in cognitive psychology called *chunks*. According to Tulving and Craik [17], a chunk is defined as "a familiar collection of more elementary units that have been inter-associated and stored in memory repeatedly and act as a coherent, integrated group when retrieved". It has been used to define expertise in many domains since Chase and Simon's early research in chess [2]. We argue that modeling chunks is important but it hasn't been well-addressed in the current Knowledge Tracing framework. In order to identify chunks in a modern data-driven manner, we propose starting from automatic extraction of stable combinations between individual skills, or between skills and difficulty factors from huge volumes of data available from digital learning systems. We think that such *chunk units* contain different complexity levels, and more complex chunk units can be constructed from simpler chunk units, so they could and should be arranged hierarchically. So we propose a hierarchical Bayesian network which we consider a natural fit for the skill and student model, rather than alternative frameworks [1, 14, 12].

Meanwhile, complex skill knowledge modeling has been a challenge. Starting from simple variants based on traditional KT [5], more advanced models have been put forward. However, these student models use a "flat" knowledge structure, and research works that consider relationships among skills mostly focus on prerequisite relations [3] or granularity hierarchy [13]. Regarding the data-driven evaluations of student models, problem-solving performance prediction metrics [7, 5] have raised some growing concerns [6, 9]. A recent learner outcome-effort paradigm and a multifaceted evaluation framework [6, 9] offer promising methods that we plan to extend. We also plan to conduct classroom studies that deploy a new adaptive learning system that is based on our proposed student model.

2. PROPOSED CONTRIBUTIONS

The first contribution we expect to achieve is to present a novel perspective and data-driven approach for building (skill and) student models with *chunks*. Second, we aim to present a novel multifaceted data-driven evaluation framework for student models that considers practically important aspects. Third, we aim to demonstrate our proposed model's impact for real-world student learning such as helping differentiating shallow and deep learning, enabling better remediation, and ultimately promoting deep learning.

3. APPROACH AND EVALUATION

3.1 Model Construction

Our proposed student model will conduct performance predictions, dynamic knowledge estimations, and mastery decisions when deployed in a tutoring system. To save space, we only describe the major components here.

3.1.1 Representing Chunk Units

To start, we plan to use the Bayesian network (BN) framework for the final skill and student model. We call our proposed model *conjunctive knowledge modeling with hierarchical chunk units (CKM-HC)* (Figure 1).

- **The first layer** consists of basic individual skills (e.g., K_1) that capture a student's basic knowledge of a skill.
- **The intermediate layers** consist of *chunk units* (e.g., $K_{1,2}$), which can be derived from smaller units that capture deeper knowledge.
- **The last layer** consists of *Mastery* nodes (e.g., M_1) for each individual skill, which reflects the idea of granting a skill's mastery based on the knowledge levels of relevant chunk units. We now assert mastery of a skill by computing the joint probability of the required chunk units being known.

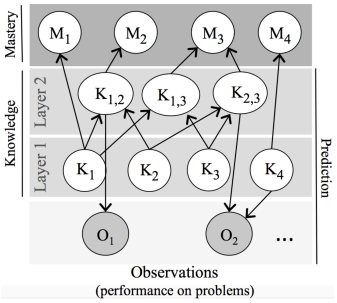

Figure 1: The BN structure of CKM-HC, with pairwise skill combinations as chunk units, in one practice time slice.

3.1.2 Identifying Chunk Units

We consider the following two frameworks to extract chunk units, with Bayesian network as the major framework:

- **Regression-based feature selection or structure learning framework.** Based on regression models, many *efficient* feature selection or structure learning methods already exist. However, the limitations of this approach include: 1) the compensatory relationship among skills is assumed; 2) it's hard to realize the evidence propagation among skills in a probabilistic way; and 3) it doesn't provide the explicit knowledge level of each individual skill. Still, we might be able to use this framework for exploratory analysis or for pre-selection, due to its potential efficiency.
- **BN-based score-and-search framework.** We can employ a search procedure for learning the structure; namely, what chunk units to include. However, if we don't limit the search space, the complexity will grow exponentially. As a result, we propose a greedy search procedural that requires a pre-ranking of the candidates for chunk units. During each iteration, it compares the cost function value of the network with a chunk unit that is newly incorporated with that of the optimal network so far.

To rank chunk units, we use the following general information that should be available across datasets or domains:

- **Frequency information based on skill to problem q-matrix.** Chunk units with higher frequencies, according to the q-matrix, can be considered to be more typical or stable patterns to be modeled.
- **Performance information based on student performance data.** We can employ various strategies, such as giving higher scores to chunk units with larger difference in the estimated difficulty between the current chunk unit and its hardest constituent skill (unit).
- **Natural language processing on the problem (solution) text.** We can consider information such as the textual proximity and semantics that can be obtained by automatic text analysis (or natural language processing).

To further improve the *interpretability*, *robustness* and *generality*, we can also use some domain-specific knowledge to extract more meaningful or typical chunk units. For example, in programming, we can use the abstract syntax tree as in [15]. However, there are still two other challenges:

- **Model run-time complexity.** Since the network involves latent variables, we use Expectation-Maximization, which computes the posteriors of latent variables in each iteration, which can be a time-consuming process.
- **Temporal learning effect.** It is also challenging to consider the temporal learning effect in such a complex network. As a first step, we ignore it during the model learning process, while maintaining the dynamic knowledge estimation during the application phase, as in [3].

We expect to explore some efficient implementations and techniques (such as re-using some posteriors or using approximate inference) to address these two challenges.

3.2 Model Evaluation

We will conduct both data-driven and classroom study evaluations to compare our model with alternatives, including traditional KT-based models [4, 5], and BN-based models with chunk units incorporated in a non-hierarchical way.

3.2.1 Data-driven Evaluation

First, we will conduct data-driven evaluations that consider:

- **Mastery accuracy and effort**. The basic idea of the mastery accuracy metric is that once a student model asserts mastery for an item's required skills, the student should be very unlikely to fail the current item. Meanwhile, the mastery effort metric empirically quantifies the number of practices that are needed to reach mastery of a set of skills. These metrics extend our approach in [6].
- **Parameter plausibility.** This metric investigates how much the fitted parameters can satisfy a model's assumptions and can be interpreted by a human. This is based on our recent Polygon evaluation framework [9].

- **Predictive accuracy of student answers.** This metric evaluates how well the new model predicts the correctness of a student's answer, or the content of a student's solution, based on the problem type.

3.2.2 *Classroom study evaluation*

We will conduct classroom studies, based on an adaptive learning system that applies our new student model. This system will contain a new open student model interface and a new recommendation engine that will be enabled by our new student model. We will focus on following questions:

1. Do students agree more with the knowledge and mastery inference obtained from the new student model?
2. Does the new student model increase students' awareness of pursuing true mastery?
3. Does the new student model enable more helpful recommendation or remediation?
4. Do students using the new adaptive learning system enabled by the new student model achieve deeper learning which is measured by specifically designed tests?

4. CURRENT WORK

We have conducted preliminary studies with skill chunk units extracted from pairwise skill combinations on a Java programming comprehension dataset and a SQL generation dataset collected across two years from University of Pittsburgh classes. Due to the runtime limitation, we employed a heuristic approach to choose skill combinations (without a complete search procedural), and conducted data-driven evaluations (by 10-fold cross validation). We found that incorporating pairwise skill combinations can significantly increase mastery accuracy and more reasonably direct students' practice efforts, compared to traditional Knowledge Tracing models and its non-hierarchical counterparts. The details of this study are reported in [10].

5. ADVICE FOR FUTURE WORK

I am seeking advice on any of the following aspects:

1. Is this idea both significant and valuable? For example, can it be connected or applied in a broad range of tutoring systems or domains?
2. Are there any datasets, domains or tutoring systems suitable for exploring this idea? What should be the desirable characteristics of the datasets?
3. Are there better representations for skill chunks within or beyond Bayesian networks (e.g., Markov random field, case-base reasoning)? Are there better techniques to identify such units?
4. Are there any suggestions for the overall procedures of this research? For example, should we do a user study to investigate this phenomenon before data mining? If so, how should we design such a study, since we can only test limited chunk units? Should we construct ideal datasets where chunk units are expected to be significant, rather than focusing on existing datasets?
5. How should we situate our definition of chunk units in a broader context considering different domains, problem (task) types and cognitive psychology theories? Is *chunk* the right word? What's its connection with production rules, declarative and procedural knowledge, Bloom's taxonomy?

6. REFERENCES

[1] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashley, and T.-W. Chan, editors, *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 164–175. Springer Berlin / Heidelberg, 2006.

[2] W. G. Chase and H. A. Simon. Perception in chess. *Cognitive psychology*, 4(1):55–81, 1973.

[3] C. Conati, A. Gertner, and K. Vanlehn. Using bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12(4):371–417, 2002.

[4] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.

[5] Y. Gong, J. E. Beck, and N. T. Heffernan. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Proc. 10th Int. Conf. Intelligent Tutoring Systems*, pages 35–44. Springer, 2010.

[6] J. P. González-Brenes and Y. Huang. Your model is predictive but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In *Proc. 8th Intl. Conf. Educational Data Mining*, pages 187–194, 2015.

[7] J. P. González-Brenes, Y. Huang, and P. Brusilovsky. General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge. In *Proc. 7th Int. Conf. Educational Data Mining*, pages 84–91, 2014.

[8] N. T. Heffernan and K. R. Koedinger. The composition effect in symbolizing: The role of symbol production vs. text comprehension. In *Proc. 19th Annual Conf. Cognitive Science Society*, pages 307–312.

[9] Y. Huang, J. P. González-Brenes, R. Kumar, and P. Brusilovsky. A framework for multifaceted evaluation of student models. In *Proc. 8th Int. Conf. Educational Data Mining*, pages 203–210, 2015.

[10] Y. Huang, J. Guerra, and P. Brusilovsky. Modeling skill combination patterns for deeper knowledge tracing. In *the 6th Int. Workshop on Personalization Approaches in Learning Environments (In Submission)*, 2016.

[11] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012.

[12] B. Mostafavi and T. Barnes. Evolution of an intelligent deductive logic tutor using data-driven elements. *International Journal of Artificial Intelligence in Education*, pages 1–32, 2016.

[13] Z. A. Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan. The effect of model granularity on student performance prediction using bayesian networks. In *Proc. 11th Int. Conf. User Modeling*, pages 435–439. Springer, 2007.

[14] R. Pelánek et al. Application of time decay functions and the elo system in student modeling. *Proc. 7th Int. Conf. Educational Data Mining*, pages 21–27, 2014.

[15] K. Rivers and K. R. Koedinger. Data-driven hint generation in vast solution spaces: a self-improving python programming tutor. *International Journal of Artificial Intelligence in Education*, pages 1–28, 2015.

[16] E. Soloway and K. Ehrlich. Empirical studies of programming knowledge. *IEEE Trans. Software Engineering*, SE-10(5):595–609, 1984.

[17] E. Tulving and F. I. Craik. *The Oxford handbook of memory*. Oxford: Oxford University Press, 2000.

Using Case-Based Reasoning to Automatically Generate High-Quality Feedback for Programming Exercises

Angelo Kyrilov
University of California, Merced
5200 North Lake Road
Merced, CA 95343, USA
akyrilov@ucmerced.edu

ABSTRACT

My research explores methods for automatic generation of high-quality feedback for computer programming exercises. This work is motivated by problems with current automated assessment systems, which usually provide binary ("Correct"/"Incorrect") feedback on programming exercises. Binary feedback is not conducive to student learning, and has also been linked to undesirable consequences, such as plagiarism and disengagement.

We propose a Case-Based Reasoning approach to utilize knowledge created by human instructors in order to automatically generate comparable responses for students that submit incorrect solutions to programming exercises. Such a system would offer significant labor savings for instructors, without sacrificing the quality of student learning.

Preliminary experiments have demonstrated the strength of our Case-Based Reasoning approach and its potential impact, especially in MOOCs. Further research is being conducted in order to refine the procedure and to evaluate it effect on student learning.

1. INTRODUCTION

Computer programming is becoming an essential skill in today's economic climate. This has led to significant enrollment increases for introductory Computer Science (CS) courses, as students from virtually all disciplines are required to learn programming. In order to cope with the increased workload, many CS educators rely on automated assessment systems for programming exercises.

Automated Assessment systems for computer programming exercises have been studied widely. [1] provides an overview of automated assessment approaches, and [7] studied the effectiveness of automated assessment on student learning. The authors found that systems which offer instant feedback and allow for multiple resubmissions are helping students to learn.

Some researchers are opposed to using such systems, mainly because of the poor quality of feedback they offer students. In many cases feedback is limited to a binary response ("Correct"/"Incorrect"). [2, 6] argue that in order for learning to take place, students who have generated incorrect solutions to a particular programming exercise, need to be given *guidance* by an expert programmer, and that simply pointing out the presence of an error is not enough.

We studied the effects of binary feedback on students and found that it increases their propensity to cheat on programming assignments and/or disengage from the course material [4]. A possible explanation for this is that since a binary response does not explain the reasons for failure, nor does it suggest a possible strategy to resolve the problem, students are often left with little choice but to cheat or given up on the exercise.

In [3], we proposed a Case-Based Reasoning approach to address the issues surrounding binary instant feedback. The idea is to use knowledge previously generated by human instructors in order to automatically build meaningful responses to incorrect programs submitted by students. In practice, such a system would have a signifiant impact in both traditional classroom environments as well as Massive Online Open Courses (MOOCs). We believe that automated feedback, comparable in quality to human-generated responses, will address motivation problems in MOOCs, which is expected to lead to increased completion rates. In regular university settings, the labor savings will allow instructors and teaching assistants to spend more time on activities beneficial to their students, rather than grading or debugging students' code.

The rest of the paper is organized as follows. Section 2.1 is an overview of our research, and a motivation for the chosen directions. Section 3 presents preliminary results, and highlights the potential contributions of this work. Section 4 outlines research questions that will be explored in future studies, and Section 5 contains concluding remarks.

2. RESEARCH TOPIC

2.1 Case-Based Reasoning

Case-Based Reasoning (CBR), first introduced by Schank [5], is a problem solving framework that uses past experiences to solve problems. Past experiences, referred to as a *cases*, are stored in a database, known as the *case base*. A single case consists of a problem description and a solution.

When a new problem, or a *query*, is encountered, the CBR system retrieves past cases whose problem descriptions are similar to the new problem, and uses the past solutions to generate instructions on how to solve the query. If executing the instructions does not lead to a solution of the problem, then the instructions are revised and evaluated again. Revisions may take place multiple times, until the solution generated by the system is accepted. At this point a new case, made up of the query and the accepted solution, is stored in the case base, making additional knowledge available for future queries. Due to its ability to create new knowledge in this way, CBR is considered a machine learning technique.

The CBR process can be summarized as the following four stages, illustrated graphically in figure 1:

1. **Retrieve:** Retrieve past cases that are similar to the query.
2. **Reuse:** The retrieved cases are used to generate a solution to the query.
3. **Revise:** The solution generated in the last step is evaluated and modified if necessary.
4. **Retain:** A new case, made up of the query and the solution are stored in the case base.

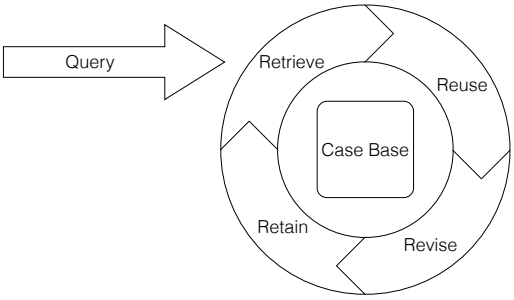

Figure 1: The case-based reasoning methodology

2.2 Proposed System

The automated assessment system we propose utilizes a Case-Based Reasoning approach to automatically assess computer programming exercises and provide feedback to students. We define a case to be a pair made of an incorrect computer program P, and instructor-generated feedback F. A computer program is deemed incorrect if it does not produce the expected outputs for a given programming exercise. Cases are therefore exercise-specific. Our case base is simply a collection of such cases.

For the retrieval stage, we need to define method of computing similarity between cases. Two cases (P_1, F_1), and (P_2, F_2) are said to be similar if P_1 is *similarly incorrect* to P_2. Two programs are similarly incorrect if they both contain the same bugs, therefore corrective feedback for one of the programs is equally appropriate for the other. In the reuse stage, we use the feedback retrieved at the previous step, without any modifications. This is possible due to the way we have defined the similarity metric for two cases.

The revise step, if necessary, will be performed by a human instructor. This is the way the system creates new knowledge. The revise procedure will be invoked if the student repeatedly submits an incorrect solution to the same exercise. This would suggest that the feedback offered by the system has not been helpful to the student. Once a correct solution has been submitted by the student, a new case is stored in the database. This case is made up of the original incorrect source code and the feedback that led to the submission of a correct solution. Figure 2 is a graphical representation of the proposed system.

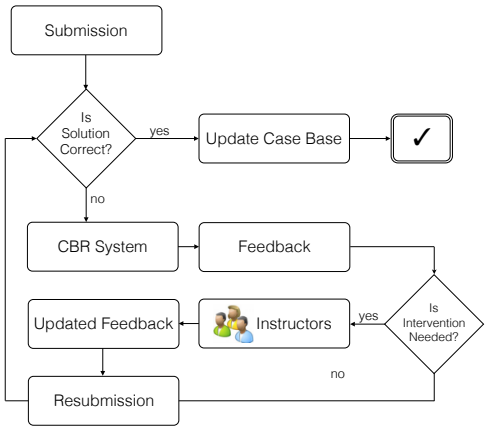

Figure 2: A flowchart of proposed system

2.3 Motivation

Previous research on Case-Based Reasoning has shown that the technique is most effective when similar problems are encountered often and when similar problems have similar solutions. Both of these conditions hold in the context of computer programming exercises. Indeed, CS educators often see the same mistake made by many students, and due to the asynchronous nature of laboratory sessions, the instructor is forced to give the same explanation to multiple students. The second condition, that similar problems have similar solutions holds true as well. If two or more students have all made the same mistake, they will all benefit from the same explanation. There could be multiple ways to explain the same mistake, and some students may find one explanation more beneficial than others. This is easily addressed by allowing the system to store multiple feedback comments per case, and present them sequentially upon unsuccessful attempts. The system can also keep track of the likelihood a particular feedback comment will lead to a successful resubmission and use this information to determine the order in which comments will be presented. Both conditions have been verified experimentally, with results presented in Section 3.

3. PRELIMINARY RESULTS

To test the soundness of our proposed system, we gathered student submissions from an undergraduate Computer Science course where students were required to complete pro-

gramming exercises on a weekly basis. Students uploaded their solutions to an automated assessment system that evaluated their correctness using unit testing.

The first research question we sought to answer was whether or not our proposed system was feasible. We randomly selected 5 exercises and extracted all the incorrect submissions for each one. We then manually clustered them according to their incorrectness. The results from this clustering procedure are presented in Table 1.

Exercise Number	Incorrect Submissions	Cluster Count	Largest Cluster	Smallest Cluster
1	111	4	54	2
2	82	10	18	1
3	73	11	19	1
4	28	8	15	1
5	26	8	13	1

Table 1: Summary of clustering experiment

It is clear from Table 1 that the same mistakes are made by many different students. This is indicated by the large values in the "Largest Cluster" column. In 4 of the 5 exercises we considered, there were mistakes committed by only one student, but small clusters are generally rare.

A more interesting and significant finding was that the number of clusters is relatively small compared to the total number of incorrect submissions. The average number of clusters is 8. This means that there are only 8 different mistakes that students are making, on average. This result is significant because an instructor with an empty case base will only need to grade 8 exercises by hand. The CBR system would be able to provide the appropriate feedback to every subsequent incorrect submission. The number of clusters is also not expected to grow with the number of students enrolled in the class. This is because the number of clusters is a function of the problem, not the number of students.

If the system scales well, it would enable MOOC instructors to provide corrective feedback to tens of thousands of students who have submitted incorrect solutions to programming exercises. This is likely to increase student engagement with the material and improve overall completion rates.

4. FUTURE WORK

In order to realize our system design, we need a reliable way to automatically detect similarity with respect to incorrectness between two programs. Our initial approach was to compute this similarity based on the unit tests. That is if two programs fail the exact same set of unit tests then they are deemed to be similarly incorrect. This is a reasonable first approach but it generates many false positives and false negatives. To ensure true scalability, the false matches need to be kept to a minimum. Methods involving static analysis of source code will likely need to be employed.

Further investigation of our scalability claims is also needed. More submission data would have to be analyzed and relationships between class size and number of clusters would need to be formally established.

Once the system has been completed, it should be deployed in a classroom and its effectiveness should be studied.

5. CONCLUSION

My research is focused on improving the quality of instant feedback generated by automated assessment systems for programming exercises. Many instructors are using automatic grading systems that are limited to providing binary feedback, which has been shown to hinder student learning and lead to plagiarism and disengagement.

We propose a Case-Based Reasoning approach to designing an automated assessment system for programming exercises capable of instantly delivering high-quality feedback, comparable to guidance a human instructor might provide to a struggling student. The system uses feedback previously generated by human instructors and delivers it to students who make similar mistakes to ones seen before.

This is an effective technique since the same mistakes are made by many different students and there are relatively few distinct mistakes. This translates into significant labor savings for instructors and teaching assistants. With our system in place, an instructor will only have to address a specific problem once. All subsequent occurrences will be handled automatically by the system.

Further research is currently being conducted on finding a reliable metric for similarity with respect to incorrectness of computer programs. Several static analysis techniques are being explored. Attempts are also being made to formalize relationships between the class size and the number of unique errors that can be made on an exercise. We postulate that for reasonably sized programming exercises, the number of unique errors will stay low even in MOOC environments where class sizes can be in the hundreds of thousands.

6. REFERENCES

[1] K. M. Ala-Mutka. A survey of automated assessment approaches for programming assignments. *Computer science education*, 15(2):83–102, 2005.

[2] T. Beaubouef and J. Mason. Why the high attrition rate for computer science students: Some thoughts and observations. *SIGCSE Bull.*, 37(2):103–106, June 2005.

[3] A. Kyrilov and D. C. Noelle. Using case-based reasoning to improve the quality of feedback provided by automated grading systems. In *Proceedings of the International Conference on E-Learning*, pages 384–388, 2014.

[4] A. Kyrilov and D. C. Noelle. Binary instant feedback on programming exercises can reduce student engagement and promote cheating. In *Proceedings of the 15th Koli Calling Conference on Computing Education Research*, Koli Calling '15, pages 122–126, New York, NY, USA, 2015. ACM.

[5] R. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, New York, NY, USA, 1982.

[6] G. N. Walker. Experimentation in the computer programming lab. *Inroads*, 36(4):69–72, 2004.

[7] D. Woit and D. Mason. Effectiveness of online assessment. *SIGCSE Bull.*, 35(1):137–141, Jan. 2003.

Predicting Off-task Behaviors in an Adaptive Vocabulary Learning System

SungJin Nam
School of Information
University of Michigan
Ann Arbor, MI 48109
sjnam@umich.edu

ABSTRACT

In many studies, engagement has been considered as an important aspect of effective learning. Retaining student engagement is thus an important goal in intelligent tutoring systems (ITS). My current studies with collaborators on Dynamic Support of Contextual Vocabulary Acquisition for Reading (DSCoVAR) include building prediction models for students' off-task behaviors. By extracting linguistically meaningful features and historical context information from interaction log data, these studies illustrate how some types of off-task behavior can be modeled from behavioral logs. The results of this research contribute to existing studies by providing examples of how to extract behavioral measures and predict off-task behaviors within a vocabulary learning system. Identifying off-task behaviors can improve students' learning by providing personalized learning materials: for example, off-task behavior classifiers can be used to achieve more accurate predictions of the student's vocabulary mastery level, which in turn can improve the system's adaptive performance. Toward our goal of developing highly effective personalized vocabulary learning systems, this research would benefit from expert feedback on issues that include: principled approaches for adaptive assessment and feedback in a vocabulary learning system; and alternative methods for defining and generating off-task labels.

Keywords

Engagement, off-task behaviors, prediction model, log data, intelligent tutoring system, adaptive system

1. INTRODUCTION

Engagement has long been considered as an important aspect of learning [17, 16]. Engagement is a comprehensive behavior that reflects an integration of different aspects of a person's cognitive state [11, 6, 7]. A student's engagement level while using the system can vary with time, and it can be influenced by many factors, such as the difficulty of questions, prior experience with similar technology, and individual interests or motivation [14, 1]. Thus, measures related to engagement need to consider the multidimensional construct of engagement and clarify which types of engagement are going to be measured in the study [18].

Other studies based on digital learning environments tend to capture engagement based on behavioral signals. Studies on intelligent tutoring systems (ITS) often used features like response time, number of erroneous attempts, and frequent accessing of hint messages to predict students' engagement [2, 4]. Studies in Massive Online Open Courses (MOOC) included features like the number of lecture videos seen, participation in pop-up quizzes, and social interactions like frequency of article posting or comments in the discussion forum, to predict the student's overall participation level [10, 15]. These studies showed that data traces of observable behavior can be used to predict student engagement, often operationalized as a classroom attitude observed from instructors or a survival rate of enlisted courses in a MOOC.

The purpose of this research topic is to model a particular subset of students' off-task behaviors while they use a vocabulary learning system, based on observations of their interaction from log data. In our study, each student response to an assessment question posed by the system was defined as an off-task behavior if it contained less serious, patterned, or repetitive errors [13, 12]. Key research questions on this topic that I will explore include: (1) identifying important predictive features of off-task behaviors in vocabulary learning systems that can be collected from log data, (2) evaluating different modeling methods that can help to develop more accurate prediction models for off-task behaviors, and (3) suggesting effective adaptive strategies for vocabulary learning systems that will help to sustain student's engagement and thus improve their learning outcomes and experience. The results from our current studies are expected to be used maximize the efficiency and long-term effectiveness of student learning outcomes.

2. CURRENT WORK AND RESULTS

Currently, I am working on developing a contextual word learning (CWL) system called Dynamic Support of Contextual Vocabulary Acquisition for Reading (DSCoVAR)[1]. DSCoVAR is an online vocabulary learning system that teaches K-12 students how to figure out the meaning of a word they don't know (sometimes called the *target word*) by using clues from the target word's surrounding context[8].

The DSCoVAR curriculum consists of three sessions: pre-

[1]http://dscovar.org

test, training, and post-test sessions. Questions in the pre- and post-test sessions include multiple types of questions measuring the student's knowledge on vocabulary before and after the training session. The training session consists of an instructional video and practice questions that teach the student different strategies for figuring out the meaning of an unknown target word by using clues from nearby words in the surrounding sentence. Students learned, and were tested on, a family of words known as Tier-2 words, which are words that are critical for understanding more advanced texts, but that are relatively rare in everyday use. These target words were expected to be difficult, but at least familiar or known to a small number of students. (In our first experiment, participants reported that they were Familiar with 26% of the Tier 2 target words, followed by 21% Known, and 53% Unknown (N=33) [13].)

2.1 Feature Extraction

In previous studies [13, 12], we analyzed students' responses in the pretest session and developed prediction models for off-task behaviors based on behavioral features extracted from log data. During sessions, DSCoVAR recorded how students interacted with the system by storing time-stamped event data and students' text responses. Based on the collected log data, we extracted two types of variables: response-time variables (RTV) and context-based variables (CTV). These variables contain more meaningful student behavior information than the raw log data, and are used as predictor variables in our off-task behavior classifiers.

RTVs collect information right after the student submits his or her response for each question, including time spent to initiate and finish typing a response, the number of spelling and response formatting errors, and orthographic and semantic similarity between the response and the target word. CTVs include history-based measures relating to how the student performed in previous trials (with different window sizes of 1, 3, 5, and 7), such as the average proportion of off-task responses in previous trials and average orthographic or semantic overlap between the current response and previous responses. Lastly, human raters created labels for off-task behaviors from log data. By using criteria based on Baker et al. [3], we obtain labels for certain types of off-task behavior, i.e. when responses seemed less serious and patterned, or when they involved repetitive errors.

2.2 Modeling Off-task Behaviors

With the RTVs and CTV features described above, we build off-task prediction models via mixed effect models and structure learning algorithms. Mixed effect models, such as the generalized linear mixed effect model (GLMM) or hierarchical Bayesian model, are suitable for analyzing the log data from ITS since they can account for variance across repeated measures like multiple responses from a single student or a particular target word.

Table 1 and 2 show the results of the GLMM model learned by the stepwise algorithm for predicting the off-task labels from RTV and CTV variables. GLMM includes random intercepts for target words and students, and the effect of random slopes for the student's prior familiarity level to the target word mentioned above 2 [13, 12]. The results show that RTV features like response length and orthographic similarity between the response and the target word are statistically significant for explaining the specific types of off-task behavior that we identified for the study. CTVs like average semantic similarity between the current response and previous three responses and orthographic similarities with previous seven responses were also significant. This model showed a better area under the curve statistic from ROC curve (0.970) than the RTV-only GLMM model (0.918).

Table 1: GLMM results for fixed effect variables (all predictors are statistically significant ($p < 0.001$))

Variables	Coeff	SE	z
(Intercept)	0.50	0.62	0.82
RTV: Response Length	-0.22	0.05	-4.10
RTV: Ort. Similarity	-5.98	1.79	-3.34
CTV: Sem. Similarity (prev. 3)	0.11	0.03	4.35
CTV: Ort. Similarity (prev. 7)	11.4	1.81	6.33

Table 2: GLMM results for random effect variables

Variables	Var.	Corr.
Target (Intercept)	1.05	
Target-Unknown:Known	2.47	-1.00
Target-Unknown:Familiar	23.0	-1.00
Subject (Intercept)	3.67	

Structure learning algorithms, such as the stepwise regression and the Hill-climbing algorithm, were used for automatically learning the model structure of off-task prediction models. The stepwise algorithm was useful in selecting which variables can bring the better fit to the regression model based on criteria like AIC or BIC. The Hill-climbing algorithm was helpful for identifying the complex interaction structures between variables based on conditional probabilities. By combining findings from different structure learning algorithms, we confirmed that adding interaction structures is helpful for prediction, especially with RTV-only models. An example of interaction structures learned from the Hill-climbing algorithm is shown in Figure 1.

3. PROPOSED CONTRIBUTIONS

First, the current work contributes to existing ITS studies by suggesting methods for extracting meaningful information from log data. For example, RTVs provided meaningful information to understand student performance on specific questions by using various language processing techniques, such as orthographic similarities measured using character trigrams, and semantic similarities measured using Markov Estimation of Semantic Association [9]. CTVs provided information on historical patterns of off-task behaviors. Combined with mixed effect models, our results suggest that traditional predictive features, such as time spent for initiating and finishing the response or number of error messages, can be substituted (when available) with features based on variance in repeated measures and contextual information.

Second, identifying off-task status at the item level can be a starting point for managing student engagement systematically, by letting the learning system know when to intervene in helping the student regain their engagement to the task. Off-task classifiers in the current studies provided examples of automatized models for checking student engagement in a vocabulary learning system.

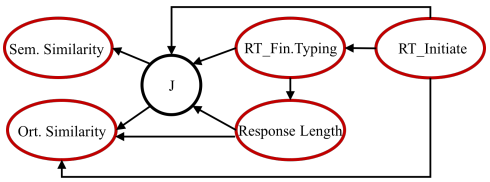

Figure 1: Interaction structure of RTVs learned by the Hill-climbing algorithm (Node J: Off-task label)

Third, this research can be helpful for achieving more accurate predictions on the student's vocabulary mastery level. For example, suggested classifiers provide item-level prediction for off-task behaviors based on previous responses. These results can be helpful for distinguishing between intentionally missed questions and accidentally erroneous responses, which in turn can be used to improve estimates provided by existing student learning prediction models, such as item response theory [5].

4. FUTURE DIRECTIONS

A key goal of this research is to build an adaptive vocabulary learning system. By using results from our current studies, we will implement an initial adaptive mechanism in DSCoVAR that personalizes the difficulty of training session's questions based on a student's estimated vocabulary mastery. This approach is expected to help retain student engagement with the system by providing the right level of 'desirable difficulty' while also making more efficient use of the student's learning time. However, it is unclear how features related to perceived question difficulty, such as amount of information given from feedback messages or size of spacing between questions that share the same target, could be used to model the overall student engagement with the question. Advice from experienced researchers on adaptively controlling task difficulty would help guide this research on personalized training to students.

Our current work depends on defining a specific type of off-task behavior, with labels generated from two human judges. While the inter-rater agreement was reasonable (Cohen's Kappa of 0.695) [12], it is an expensive process and the number of collectible judgments are limited. An alternative approach could be to use crowd-sourcing for labeling the log data. However, converting this expert labeling task into a fragmentary job for anonymous workers may require more carefully designed instructions and robust methods for validating the credibility of labels. Expert guidance on alternate definitions of off-task behavior, and improved approaches for gathering larger amounts of labeled data based on these definitions, would be helpful for expanding future studies.

5. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Educational Sciences (IES R305A140647) through a grant to the University of Michigan. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. I also thank Dr. Kevyn Collins-Thompson and Dr. Gwen Frishkoff for their guidance and suggestions.

6. REFERENCES

[1] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, 2014.

[2] I. Arroyo and B. P. Woolf. Inferring learning and attitudes from a bayesian network of log file data. In *AIED*, pages 33–40, 2005.

[3] R. S. Baker, A. T. Corbett, and K. R. Koedinger. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems*, pages 531–540. Springer, 2004.

[4] J. E. Beck. Engagement tracing: using response times to model student disengagement. *Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, 125:88, 2005.

[5] S. E. Embretson and S. P. Reise. *Item response theory for psychologists*. Psychology Press, 2000.

[6] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109, 2004.

[7] J. A. Fredricks and W. McColskey. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement*, pages 763–782. Springer, 2012.

[8] G. Frishkoff, K. Collins-Thompson, and S. Nam. Dynamic support of contextual vocabulary acquisition for reading: An intelligent tutoring system for contextual word learning. In *Adaptive Educational Technologies for Literacy Instruction*. Taylor & Francis, Routledge:NY, In Press.

[9] G. A. Frishkoff, K. Collins-Thompson, C. A. Perfetti, and J. Callan. Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment. *Behavior Research Methods*, 40(4):907–925, 2008.

[10] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 170–179. ACM, 2013.

[11] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[12] S. Nam, K. Collins-Thompson, and G. Frishkoff. Modeling real-time performance on a meaning-generation task. In *Annual Meeting of the American Educational Research Association*. AERA, 2016.

[13] S. Nam, K. Collins-Thompson, G. Frishkoff, and L. Hodges. Measuring real-time student engagement in contextual word learning. In *The 22nd Annual Meeting of the Society for the Scientific Study of Reading*. SSSR, 2015. `https://goo.gl/CvTL1K`.

[14] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008.

[15] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[16] B. Ravindran, B. A. Greene, and T. K. Debacker. Predicting preservice teachers' cognitive engagement with goals and epistemological beliefs. *The Journal of Educational Research*, 98(4):222–233, 2005.

[17] J. P. Rowe, L. R. Shores, B. W. Mott, and J. C. Lester. Integrating learning and engagement in narrative-centered learning environments. In *Intelligent Tutoring Systems*, pages 166–177. Springer, 2010.

[18] G. M. Sinatra, B. C. Heddy, and D. Lombardi. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13, 2015.

Estimation of prerequisite skills model from large scale assessment data using semantic data mining

Bruno Elias Penteado
ICMC - Institute of Mathematics and Computer Science
University of Sao Paulo
Av. Trabalhador São-Carlense, 400, São Carlos, Brazil
brunopenteado@usp.br

ABSTRACT

Learning sequences are important aspects in learning environments. Students should learn by moving gradually from simpler to more complex concepts, promoting deeper levels of learning. This feature is usually embedded in most intelligent learning environments to guide the student in the study of subject matter. The organization of this knowledge structure is usually an intensive effort of human experts, in creating a logical ordering of what is to be taught - determining the concepts and the prerequisite relations among them. In recent years, some methods have been developed for dealing with this knowledge structuring using data coming from logs of learning environments, applying data mining techniques to discover prerequisite rules and create directed graphs of prerequisites. These methods model both assessment items and skills underlying those items. The automatic methods developed so far present a semantic gap between the probabilistic analysis and the expert knowledge, sometimes causing confusion with the results. This research aims to bridge this gap by adding a minimal layer of semantic information to help in the data mining process. As an application, we intend to analyze large-scale assessment datasets, considering its specificities, and evaluate if those hybrid models can improve the prediction of item success.

Keywords

Skill model, knowledge structure, data mining, semantic data mining.

1. INTRODUCTION

Skills prerequisite structure is an important component in *domain modeling*, used in intelligent learning environments and which serve as a basis for planning learning sequences and adaptive strategies for tutoring systems. Analogously, most intelligent learning environments uses a *student model* for the automatic adaptation of teaching strategies and as a overlay of domain model, influencing how the automatic intervention is carried out. Human experts usually define such prerequisite structure; however, they are rarely validated empirically and improved for better results.

For most of the large scale assessments, the current approach considers all knowledge in a single unidimensional scale, which considers the item difficulty in its ordination. *Computer adaptive tests* tend to use predominantly this ordination for item selection in diagnostic assessments. This approach raises some issues: the *interpretability* of results, since a single value is used to represent a knowledge in a large domain; and the *agreement* about the structure, since most experts cannot see a direct, unidimensional relationship among skills. Given the amplitude of skills, experts seem to agree on other sorts of dependencies, not just the simple ordination for item difficulty. For instance, in the field of Physics, an easy item of spatial movement might not be considered as a prerequisite for a difficult item in geometric optics, since they belong to different branches.

On the other hand, the process of manual creation of these dependencies is highly costly, time-consuming and presents large disagreement among experts modeling the same domain. Pavlik et al. [1] point to 3 other factors: the description of irrelevant skills, redundancy among skills and the ordination of those skills

There seems to be a semantic gap between the automatic extraction from data and the mapping made by human experts. This research aims to explore this gap, trying to bridge it using semantic data mining, and combining the advantages of both approaches.

2. PREVIOUS WORK

The process of prerequisite structure derivation from observable variables (such as assessment items) from data has been investigated by many researchers; yet, the skill modeling is still an open issue, since a student's knowledge is a latent variable, not being observed directly. In [2] it is proposed the POKS (Partial Order Knowledge Structure) algorithm to learn the dependency structure among items, composed only by the observable nodes (answers to the items), outperforming Bayesian networks algorithm, both in predictive performance and computational efficiency. In [1] POKS algorithm is applied to analyze the relations among skills, using observable items and use the result to cluster redundant skills, with a high degree of covariance, simplifying the domain model and determining it structure. In [3] a method is proposed to determine dependency relations among curricular units from student's performance data, using a binomial test for every pair of skills, to evaluate the existence of a prerequisite relationship between them. In [4] a frequent association rules mining method is proposed to discover concept maps, but not considering the uncertainty in the process of knowledge transfer of the student to his performance. In [5] the structure is derived from noisy observations using log likelihood calculated between the precondition model and the model in which the skills are all independent on each pair of skills to estimate which model better fits the student's data. In [6] causal discovery algorithms are used to find a skill prerequisite structure applying statistical tests in the latent variables. In [7] is proposed a probabilistic association rules mining method, having the probabilistic knowledge states estimated by an evidence model, to find a structure from performance data.

In semantic technologies, ontologies are explicit specifications of conceptualization and a formal way to define the semantics of knowledge and data. Dou et al. [8] surveys this semantic data mining in multiple domains - formal ontologies have been introduced to semantic data mining to: i) bridge the semantic gap between data, data mining algorithms and results; ii) provide data mining algorithms with a priori knowledge, guiding the mining process or reducing the search space; iii) provide a formal way for representing the data mining flow, from data preprocessing to mining results. Bellandi et al. [9] presented an ontology-based association rule mining method, using the ontology to filter instances in the process, constraining the search space of itemsets, excluding items and characterizing others according to an abstraction level, enabling generalization of an item to a concept of the ontology. Marinica and Guillet [10] presented a post-processing method for the results of the association mining, pruning invalid or inconsistent association rules with the help of the ontology.

Large scale assessments present some specificities: they are very strict in their skill model, with reference matrices specifying what is expected in the test; they are periodic, meaning that they are applied, in some cases, in an annual basis, with no single item in common between applications; the test items are organized in blocks (incomplete balanced blocks) and the test is comprised of a few blocks with a fixed number of items, so that many versions of the test are available at a time; the items are all pre-tested before the actual application, to estimate psychometric parameters (following Item Response Theory principles) being equalized into the same scale. A challenge for this research is to work with datasets from multiple years (i.e., no common items), balanced in blocks trying to discover generalizations in the underlying skill model.

3. METHOD AND MATERIAL

In this work, we will work with microdata from ENEM – an annual Brazilian exam for high school students, used as a classification ranking for admission in many public federal universities in Brazil. This exam is composed by 4 knowledge areas (Mathematics, Natural Sciences, Human Sciences and Languages), each composed by 30 skills in the reference matrix specified for this exam. Each item is mapped to a single skill and a score is given for each of these knowledge areas. The test is composed by 45 multiple-choice items for each knowledge area, along with an essay, in a 2-day time span. Different tests are organized in an *incomplete balanced blocks* design. In this approach, each test is composed by multiple blocks of items, with fixed ordination and in increasing order of difficulty. The blocks are arranged in different tests so to alleviate possible biases like the position of an item and a fatigue factor for items in the end of the test.

The datasets contain every alternative selected by every student whom participated in the exam. We plan to conduct this study using the Mathematics dataset, from 2009 to 2014, in a sum of 270 items answered my tens of millions of students.

Working along with Math experts, we will try to create simple ontologies, just with constraints of what should or not be considered in the final model, to prune some of the spurious results.

This research will adopt a quantitative approach and use data mining techniques as a method to construct the mapping of the prerequisite structure which, from the items mapped to their respective skills and the performance data (correct and incorrect answers) for every respondent, is able to extract relations among the skills, generalized by different observations in different items.

The evaluation of the method will be based on the capacity of prediction of success on the items individually, assessing the goodness of fit against the human experts mapping. The method will be compared to state-of-the-art algorithms such as POKS, probabilistic association rules mining and with some expert mapping.

4. PRELIMINARY WORK

This is a research project in its earlier stages, narrowing the research questions to be pursued. As an initial effort, I found that more simplistic approaches tend to model just the difficulty of items in the creation of a prerequisite structure, i.e., an easier item is a prerequisite for a more difficult item, disregarding contextual information on the respective topics.

Early examples for ENEM using data from Mathematics test applied in 2014 are depicted in Figure 1 (skill prerequisites). They were generated by the author using the POKS algorithm, with source code available in [11] and show the algorithm results.

In Figure 1, the previous items were mapped to their respective skills and the algorithm was run. Skills are numbered according to the official codes available at ENEM website. We can see that some skills are more fundamental, specially numbers 1, 3, 4 and 17. Skills 12, 15 and 22 were not assessed in this test.

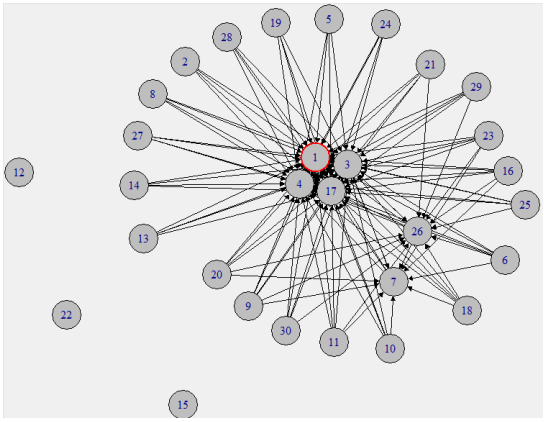

Figure 1. Prerequisite skills derived from Math assessment.

We hope, by the end of this research, discover possible prerequisite relations among skills that constitute the ENEM exam, complementing the traditional model of ordination by item difficulty in the IRT model, by creating a generalized graph of dependencies among skills, estimated from empirical data of application and combined with ontology constraints.

From this mapping, it should be possible to build an intelligent learning environment that might diagnose in which point of the graph the student is and the possible sequences he can choose to study. Another practical implication may be the interpretation of results and extension to practices in public policies. As this sort of exam is applied in different moments in K-12, the model could generalize and describe how learning happens in public education system, since literacy through high school.

5. ADVICES SOUGHT

For this doctoral consortium, advice is sought regarding some concerns:

a) *What data mining methods should be used to model these prerequisite skills?* At first, POKS was used but other methods could also be evaluated, like LFA, Rule Space and BKT. As this is a high stake exam, the skills are wider, different from other more granular skill models from ITS domains. An example (skill 17, a basic skill from Figure 1): "analyze information involving variations in quantity as a resource for argument construction". In addition, the same skill can vary a lot depending on the items being assessed. Second, items being that different and having different difficulty parameter,

b) *Should difficulty be embedded in the model?* so that different items of a same skill can influence differently in the model.

c) *Should these information be included in the model?* which may result in different graphs for different populations. Besides the standard item accuracy prediction. This dataset has no other interaction data, as in ITS systems, but has contextual data about the respondents, with high impact features in performance, like geographic region and socioeconomic status.

d) *Is it valid to measure a interrater agreement metric (like Kappa) to compare the generated model with those from experts?* as a means of comparing how close the model fit the expert modeling.

6. REFERENCES

[1] Pavlik Jr., P.I., Cen, H., Wu, L., Koedinger, K.R.: Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor. In Proceedings of the 1st International Conference on Educational Data Mining, Montreal, Canada, 77-86, 2008.

[2] Desmarais, M.C., Meshkinfam, P., Gagnon, M.: Learned Student Models with Item to Item Knowledge Structures. User Modeling and User-adapted Interaction, 16(5), 403-434, 2006.

[3] Vuong, A., Nixon, T., Towle, B.: A Method for Finding Vuong, A., Nixon, T., Towle, B.: A Method for Finding Prerequisites within a Curriculum. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands, 211-216, 2011.

[4] Tseng, S.S., Sue, P.C., Su, J.M., Weng, J.F., Tsai, W.N.: A New Approach for Constructing the Concept Map. Computers & Education, 49(3), 691-707, 2007.

[5] Brunskill, E.: Estimating Prerequisite Structure from Noisy Data. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands, 217-222, 2011.

[6] Scheines, R., Silver, E., Goldin, I.: Discovering Prerequisite Relationships among Knowledge Components. In Prerequisites within a Curriculum. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands, 211-216, 2011.

[7] Chen, Y., Wuillemin, P. H., Labat, J. M. Discovering prerequisite structure of skills through probabilistic association rules mining. In Proceedings of the 8th International Conference on Educational Data Mining, Montreal, Canada, 77-86, 2015.

[8] Dou, D., Wang, H., Liu, H. Semantic Data Mining: a survey of ontology-based approaches. In Proceedings of the 9th International Conference on Semantic Computing, Anaheim, USA, 244-251, 2015. DOI: 10.1109/ICOSC.2015.7050814.

[9] Bellandi, A., Furletti, B., Grossi, V., Romei, A. Ontology-driven association rule extraction: A case study. Contexts and Ontologies Representation and Reasoning, page 10, 2007.

[10] Marinica, C., Guillet, F. Knowledge-based interactive postmining of association rules using ontologies. Knowledge and Data Engineering, IEEE Transactions on, 22(6):784–797, 2010

[11] Desmarais, M., Bhatnager, S. Prerequisite skill structures in ASSISTments. Available in: https://github.com/sameerbhatnagar/POKS-skills. Accessed: March, 20th, 2016.

Designing Interactive and Personalized Concept Mapping Learning Environments

Shang Wang
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, Tempe AZ, USA
swang158@asu.edu

ABSTRACT

Concept mapping is a tool to represent interrelationships among concepts. Relevant research has consistently shown the positive impacts of concept mapping on students' meaningful learning. However, concerns have been raised that concept mapping can be time consuming and may impose a high cognitive load on students. To alleviate these concerns, research has explored facilitating concept map construction by presenting students with incomplete templates and concept map based navigational assistance on the learning material. However, it's not clear how these incomplete templates should be designed to address individual student needs and how concept map-based navigation can support students in creating concept maps and developing personalized navigation patterns. In this paper, I discuss my previous research in providing personalized scaffolding in concept mapping activities and describe plans of my research in exploring how personalized concept map scaffolding supported by navigational assistance could enhance student learning.

Keywords

Data mining, concept mapping, navigation, personalization, adaptive scaffolding, expert skeleton concept map.

1. Research Topic

Concept maps are graphical representations of knowledge structures, where labeled nodes denote concepts and links represent relationships among concepts. Concept mapping has been widely employed in educational settings to support student learning. Research has examined how concept mapping tools assist students in summarizing, relating, and organizing concepts [1][4]. However, there are limitations in using concept mapping. The main disadvantage of concept mapping is that the map construction is time-consuming and it requires some expertise to learn [3]. In addition, the complexity of the task often imposes high cognitive load and reduces student motivation [10].

Cañas and colleagues developed CmapTools, a computer-based concept mapping system, to support concept mapping by making it easier to construct and manage large representations for complex knowledge structures [6]. Although CmapTools provides a convenient platform for concept map construction, the system is independent from the learning content and students may encounter difficulties relating maps with resources and comparing linked concepts. To enhance concept maps with relevant resources, McClellan and colleagues designed a system that attaches resources like demos, homework and tutorials to the concept maps via keyword matching [11]. However, it might cause extraneous effort for students to process this additional information.

Apart from providing computer systems for concept map construction, other research canvassed the effect of providing students with incomplete templates called expert skeleton maps, within which some nodes and links were set as blanks, as a scaffolding aid [5]. Although studies show that the scaffolding had more positive effects on student learning than those who created concept maps from scratch [3], it's not clear how expert skeleton maps should be designed to provide better learning results. Questions like what concept nodes should be presented and what concept nodes should be left blank, how big should the expert skeleton map be, and should all students be given the same expert skeleton map, still remain unsolved. To address these challenges and the opportunities from the two directions discussed above, I propose a design of a personalized and interactive concept mapping learning environment that integrates a textbook with a concept mapping tool. This system will enable students to create maps directly from the textbook. Students will relate the created maps to the textbook content and the system will offer personalized scaffolding to facilitate concept map construction and meaningful learning. I also describe my plan of conducting an Amazon Mechanical Turk Study and an in-classroom study to test the system.

2. Proposed Contribution

2.1 Previous Work

Towards designing a personalized and interactive concept mapping learning environment, my prior work has examined how personalized expert skeleton maps affect student learning. More specifically, I studied the potential effects of an adaptive expert skeleton scaffold that contains concepts and relationships for which the student has demonstrated prior knowledge [7]. To create the adaptive expert skeleton maps, an expert concept map representing the knowledge structure from the chapter was first created as a foundation. I then mapped each question on the pretest to a certain part of the expert map to modify the expert skeleton map based on students' pretests scores. For example, if a student incorrectly answered question 4 as shown in Figure 1, the correct concept ("flower") was replaced with '???' and left open for the student to fill in. By presenting students with a map that contained their prior knowledge, I hypothesized that students would spend more effort on unknown concepts and be better supported in integrating new knowledge into prior existing knowledge structure, thus improving learning.

Figure 1. Modifying the expert map based on pre test answers.

To test my hypothesis, I conducted a study with 38 non-biology major students who were randomly assigned into three conditions: (1) adaptive scaffolding, (2) fixed scaffolding and (3) non-scaffolding. Students in the adaptive scaffolding condition received an expert skeleton map that contained nodes and links which they got correct in the pretest. Students in the fixed scaffolding condition also received a skeleton map. However, instead of tailoring the map to the student's prior knowledge, I presented them with maps from the adaptive scaffolding condition through yoked control. In this way, I was able to control for content across conditions. Finally, in the non-scaffolding condition, students constructed a map from scratch. Although I did not discover significant differences in learning gains between conditions, I found that different types of nodes in the template did lead to different learning gains on related concepts. To further investigate, I coded the key ideas in the expert map as being: added to the map by the student, already in the template, or not added. For the already existing concepts in the expert map, I further categorized the concepts that were adjacent to the newly added concepts as "close" and the ones which were more than one link away as "far". Results indicate that students benefit most from adding concept nodes to the map and benefiting more from the in template "close" nodes than the "far" and not added ones.

However, there were several limitations in the data collection that might have influenced the results. First, the number of graduate students and undergraduates was not balanced across conditions, and the learning differences in these two populations may have added extraneous noise to the learning gains. Another potential problem was that the expert skeleton maps students received might have been too large. While I assessed students on 9 key ideas, these ideas spanned more than 70 nodes in our expert map. The complexity of the given template might have imposed high cognitive load on students, reducing the benefits of the expert skeleton maps. What's more, the concept mapping system used in the study, the CmapTools, required students to type the words to create nodes. The system was also limited in terms of searching and comparing resources related to the concept maps.

2.2 System Design

Incorporating my finding discussed above, I present an iPad-based interactive concept mapping tool that is integrated with a digital textbook. When held in landscape mode, the screen splits into two, with the left side displaying the textbook and the right side showing the concept mapping panel. The built-in concept mapping view is directly associated with the learning material, so that students can construct concept maps directly from the words in the textbook, shown in Figure 2. An example of this process would be a student reads the textbook and find the concept "seed" that should be contained in the concept map, he can long click on the word and tap on the add concept button to add a node named "seed" to the concept mapping panel on the right. He can click on other concepts to add and delete links. This feature eliminates the tedious process of manually adding nodes and typing all the text while encouraging the cognitively beneficial processes of finding the important concepts and identifying the relations among them. Apart from that, a hyperlink between the node in the concept map and the words in textbook is created through the "click and add" action, allowing students to navigate through the textbook by clicking on the concept nodes. During their navigation, related concepts in the concept map and the text in the textbook are both highlighted, providing a visual comparison of key information. Since the concept maps are created by students themselves, the system enables students to form their own navigation patterns to assist them in locating key information in the textbook resource. The system is able to provide pre and post tests, which can be used to dynamically modify the expert skeleton map based on the student's prior knowledge. Furthermore, leveraging the hyperlinking navigation feature, the system enables students to click on the nodes in the expert skeleton to navigate directly to related pages.

I hypothesize that the system can alleviate the challenges discussed in the introduction and benefit students in different ways. It first allows students to easily construct concept maps via the "click and add" feature, which reduces the work of tediously typing words into the nodes while preserving the beneficial work of searching and identifying concepts to be added. The hyperlinking navigation provides more flexibility in comparing and finding connections between concepts that are located in different pages. Hyperlinking the expert skeleton map with the textbook enables students to click on the nodes provided in the expert skeleton map to see where these concepts are mentioned in the textbook. This would reduce the cognitive load of the template, which is a potential cause of reducing the effect of adaptive scaffolding in my previous work. What's more, providing students with expert skeleton maps that contain their prior knowledge would facilitate meaningful learning while they add new concept nodes to templates that represent their own knowledge structures. Since the concept nodes are already mastered by students, this approach also avoids potential shallow learning, which is a problem faced by many forms of computer-based instruction [8].

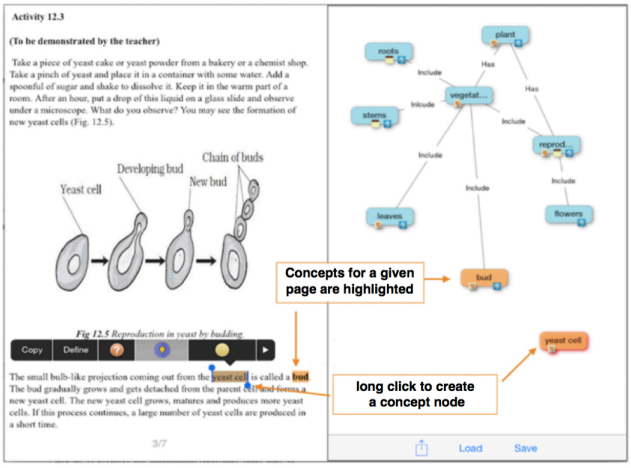

Figure 2. Interactive concept mapping system interface.

3. Results So Far

My previous study showed that types of interactions with the concept map have an effect on student learning gains [8]. However, limitations of the CmapTools and the complexity of the expert skeleton map reduced the effect of adaptive scaffolding. To solve these problems, I have implemented the proposed iPad-based concept mapping system and I'm currently running two studies to explore how different designs of expert skeleton maps and hyperlinking navigation effect learning out comes.

To test the effect of different types of scaffolding, I'm running an online study using Amazon Mechanical Turk, a human intelligent task market in which anyone can post tasks to be completed and specify prices paid for completing them. The literature indicates that Amazon Mechanical Turk could be a promising approach to get inexpensive, yet high quality data for research in psychology and social sciences [9]. However, few research has examined the

quality of Mechanical Turk data in educational studies. Thus, I plan to explore how Mechanical Turk can be used as a cost effective way to get high quality data for educational studies. In this study, participants use an online iPad simulator running the concept mapping application to construct a concept map while they learn a chapter of a high school science textbook. First, students are given a 2-minute pretest to assess prior knowledge on water pollution. Next, students are given a 3-minute training about what concept maps are and how to use the application to construct one. After the tutorial and practice, students are given a randomly modified expert skeleton map and are given 20 minutes to construct or complete the map based on the template. Finally, a posttest is given. Instead of tailoring the scaffolding specifically to student prior knowledge, I'm randomly selecting the size and concept nodes that appear in the template, in order to generate more variations of the expert skeleton map. Learning outcomes based on these different designs of expert skeleton maps could help us understand how the expert skeleton map should be designed to better facilitate learning.

Furthermore, I plan to examine how concept map-based navigation facilitates concept map construction and how it helps students to form personalized navigation patterns. I am currently working with a high school teacher to conduct a study in one of her classes, which has been using concept maps as a class activity. The study will last 20 minutes per day for 5 days and it will be a substitute for a paper-and-pencil based concept mapping activity. Students will construct the concept maps while they learn about the current textbook chapter. Students will be randomly assigned into two conditions: The hyperlinking condition, where nodes in the concept maps are hyperlinked with the textbook, and the non-hyperlinking condition. Pre and post tests will be given before and after the study. To investigate the effect of hyperlinking, I will compare the learning gains between condition. Furthermore, I plan to use data mining techniques to extract patterns within student navigation activities. For example, if a student is navigating by clicking back and forwards on two linked concept nodes, it might indicate that the student is using the textbook content to compare the concepts. If a student is navigating by clicking on a series of connected nodes, it might indicate that the student is comparing multiple concepts to understand some knowledge structure in a higher level.

4. Advice Sought

For this doctoral consortium, advice is sought regarding two major concerns. First, how should I validate the Amazon Mechanical Turk study results? I'm currently using Amazon Mechanical Turk platform for the expert skeleton map study. As I'm randomly varying the size and the concept nodes which appear in the template, I need a large number of participants to form overlaps between the student prior knowledge and the given expert skeleton map. Amazon Mechanical Turk would be a cost-efficient approach to get large amount data. However, due to the large variations in the participant population, the results from the study might not truly reveal the effect of expert skeleton map scaffolding on high school students. How could I make use of the Mechanical Turk study data to design concept mapping scaffolding to better facilitate learning?

Second, what data mining techniques can be used to analyze the hyperlinking study data? I'm interested in discovering what student behavior patterns correlate to learning outcomes and what interactions are tedious and counterproductive, and can be potentially be supported or replaced by computer technologies.

Problems discussed above are major challenges I encounter to analyze the data from the studies. Advice on these two problems will be very helpful to my work of designing personalized expert skeleton maps to facilitate concept map construction and providing hyperlinking navigation to reinforce student learning.

Acknowledgments

This research was funded by NSF CISE-IIS-1451431 EAGER: Towards Knowledge Curation and Community Building within a Postdigital Textbook.

5. REFERENCES

[1] Novak, Joseph D. "Concept mapping: A useful tool for science education."*Journal of research in science teaching* 27.10 (1990): 937-949.

[2] Azevedo, Roger, and Allyson F. Hadwin. "Scaffolding self-regulated learning and metacognition–Implications for the design of computer-based scaffolds." *Instructional Science* 33.5 (2005): 367-379.

[3] Chang, Kuo-En, Yao-Ting Sung, and S. F. Chen. "Learning through computer-based concept mapping with scaffolding aid." *Journal of Computer Assisted Learning* 17.1 (2001): 21-33.

[4] Chularut, Pasana, and Teresa K. DeBacker. "The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language." *Contemporary Educational Psychology* 29.3 (2004): 248-263.

[5] Novak, Joseph D., and Alberto J. Cañas. "The theory underlying concept maps and how to construct and use them." (2008).

[6] Cañas, Alberto J., et al. "CmapTools: A knowledge modeling and sharing environment." *Concept maps: Theory, methodology, technology. Proceedings of the first international conference on concept mapping.* Vol. 1. 2004.

[7] Wang, Shang, et al. "Personalized Expert Skeleton Scaffolding in Concept Map Construction." *Artificial Intelligence in Education*. Springer International Publishing, 2015.

[8] Burton, Richard R., and John Seely Brown. "An investigation of computer coaching for informal learning activities." *International Journal of Man-Machine Studies* 11.1 (1979): 5-24.

[9] Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?."*Perspectives on psychological science* 6.1 (2011): 3-5.

[10] Kinchin, Ian M. "If concept mapping is so helpful to learning biology, why aren't we all doing it?." *International Journal of Science Education* 23.12 (2001): 1257-1269.

[11] McClellan, James H., et al. "CNT: concept-map based navigation and discovery in a repository of learning content." *Frontiers in Education, 2004. FIE 2004. 34th Annual*. IEEE, 2004