6th International Workshop on Personalization Approaches in Learning Environments (PALE 2016)

held in conjunction with

24th ACM International Conference
User Modeling, Adaptation, and Personalization (UMAP 2016)

Halifax, Canada, July 16th, 2016

Proceedings edited by:

Milos Kravcik
Olga C. Santos
Jesus G. Boticario
Maria Bielikova

Published in CEUR workshop proceedings (Vol. 1618)
ISSN 1613-0073

http://ceur-ws.org/Vol-1618/#PALE
# Table of contents

**Preface**  
Milos Kravcik, Olga C. Santos, Jesus G. Boticario, Maria Bielikova  
1-4

**Addressing Affective States with Empathy and Growth Mindset**  
Arroyo, I., Schultz, S., Wixon, N., Muldner, K., Burleson, W., and Woolf, B. P.  
5-9

**Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario?**  
Alyuz, N., Okur, E., Okay, E., Genc, U., Aslan, S., Mete, S. E., Stanhill, D., Arnrich, B., and Esme, A. A.  
10-16

**Affective and behavioral assessment for adaptive intelligent tutoring systems**  
17-21

**Modeling Skill Combination Patterns for Deeper Knowledge Tracing**  
Huang, Y., Guerra, J., and Brasilovsky, P.  
22-30

**Exploring Requirements for an Adaptive Exercise Selection System**  
Okpo, J., Dennis, M., Masthoff, J., Smith, K., and Beacham, N.  
31-36

**Exploring the Impact of Extroversion on the Selection of Learning Materials**  
Alhathli, M., Masthoff, J., and Siddharthan, A.  
37-41

**Combining Content Analytics and Activity Tracking to Identify User Interests and Enable Knowledge Discovery**  
Vozniuk, A., Rodriguez-Triana, M. J., Holzer, A., and Gillet, D.  
42-48

**Theory-driven Learning Analytics and Open Learner Modelling: The Teacher's Toolbox of Tomorrow?**  
Kickmeier-Rust, M. and Albert, D.  
49-52

**Detecting Cheaters in MOOCs Using Item Response Theory and Learning Analytics**  
Alexandron, G., Lee, S., Chen, Z., and Pritchard, D. E.  
53-56

**WiBAF into a CMS: Personalization in Learning Environments Made Easy, pages**  
Montes García, A., Stash, N., Fabri, M., De Bra, P., Fletcher, G. H. L., and Pechenizkiy, M.  
57-62

**Towards Personal infrastructure to manage long term open learner models**  
Gilliot, J.-M., El Mawas, N., and Garlatti, S.  
63-67

**Tracking and Reacting to the Evolving Knowledge Needs of Lifelong Professional Learners**  
Ishola, O. and McCalla, G.  
68-73
ABSTRACT
Personalization approaches in learning environments are crucial to foster effective, active, efficient, and satisfactory learning. The focus of the PALE workshop series is on the different and complementary perspectives in which personalization can be addressed in learning environments (e.g., informal, workplace, lifelong, mobile, contextualized, and self-regulated learning) and offers an opportunity to present and discuss a wide spectrum of issues and solutions. In particular, this sixth edition includes 12 papers dealing with emotional engagement, affective states, personality, deep learning and complex skills, learning analytics and recommendations, educational data mining, lifelong learning, open learner models, as well as adaptive MOOCs.

CCS Concepts
• Education → Interactive learning environments • World Wide Web → Personalization • Users and interactive retrieval → Personalization.

Keywords
Personalization; Adaptive educational systems; Learning environments.

1. INTRODUCTION
The 6th International Workshop on Personalization Approaches in Learning Environments (PALE)\(^1\) took place on July 16\(^{th}\), 2016 in Halifax (Canada) and was held in conjunction with the 24\(^{th}\) ACM conference on User Modeling, Adaptation, and Personalization (UMAP 2016).

Since PALE topic can be addressed from different and complementary perspectives, PALE workshop series aims to offer a fruitful crossroad where interrelated issues can be contrasted and discussed. PALE 2016 was a follow-up of the five previous editions of PALE (which took place at UMAP 2011 – 2015) whose main contributions are compiled in the Special Issue on User Modelling to Support Personalization in Enhanced Educational Settings published by the International Journal of Artificial Intelligence in Education [1]. As a long-standing workshop series, PALE workshop has established itself as a mature channel for disseminating research ideas about personalization in learning environments. This could not be possible without the very much appreciated involvement of the program committee members (many of them supporting PALE all along these years) as well as the active participation of authors who have selected this venue to disseminate and discuss their research.

From the past experience we have identified new areas of interest in this research scope to complement the previous ones. Thus, in this workshop edition we focused on sharing and discussing the new trends in current research on how user modeling and associated artificial intelligent techniques are able to contextualize and manage the increasing amount of information coming from the task at hand and its surrounding environment in order to provide the personalization support in a wide range of learning environments, which are increasingly more sensitive to the learners and their context. This covers many interrelated fields such as: intelligent tutoring systems, learning management systems, personal learning environments, serious games, agent-based learning environments, among others.

In order to foster the sharing of knowledge and innovative ideas on these issues, PALE format follows the Learning Cafe methodology to promote discussions on open issues regarding personalization in learning environments. Four Learning Café sessions were set up for this year PALE edition. Each one consisted of brief presentations of the key questions posed by three workshop papers and subsequent small group discussions with participants randomly grouped at tables. Each table was moderated by the presenter of the paper. During the session, participants changed tables to promote sharing of ideas among the groups. The workshop ended with a summary of the discussions on each paper. In this way, participants attending the workshop could benefit both from interactive presentations, constructive work, and knowledge sharing.

In the following, we introduce PALE 2016 motivation and themes as well as present an overview of the contributions accepted and discussed in the workshop.

\(^1\) http://adenu.ia.uned.es/workshops/pale2016/
2. MOTIVATION

The target audience of the PALE workshop includes researchers, developers, and users of personalized and adaptive learning environments. Personalization is crucial to foster effective, active, efficient, and satisfactory behavior in learning situations in an increasing and varied number of contexts, which includes informal learning scenarios that are being demanded in everyday life activities and lifelong learning settings, with more control on the learner side and more sensitivity towards context. Personalization of learning environments is a long-term research area, which evolves as new technological innovations appear.

Nowadays there are new opportunities for building interoperable personalized learning solutions that consider a wider range of data coming from varied learner situations and interaction features (in terms of physiological and context sensors). However, in the current state of the art it is not clear how the new information sources are to be managed and combined in order to enhance interaction in a way that positively impacts the learning process whose nature is essentially adaptive.

In this context, suitable user modeling is needed to understand both realistic learning environments cropping up in a wider range of situations and the needs of the learners within and across them. There are new open issues in this area, which refer to detecting and effectively managing personal and context data in an increasing and varied range of learning situations in order to provide personal assistance to the learner, which can also take into account their affective state. This requires enhancing the management of an increasing number of information sources (including wearables) and big data which ultimately are to provide a better understanding of every person's learning needs within different contexts and over short-, medium-, and long-term periods of time.

This will hopefully increase learner's understanding of their own needs in terms of open learner models that are to be built from standards that support interoperability and which are to cover an extended range of available features, thus allowing for combining different external learning services as well as taking advantage of the integration of an increasing amount of information sources coming from ambient intelligence devices to gather information not only about the learner interaction, but the whole context of the learning experience. In this way, the learner modeling involves analyzing changing situations in terms of context, learners' needs and their behavior, requiring personal and collective management of the information available.

The focus of this workshop series is put on the different and complementary perspectives in which personalization can be addressed in learning environments (e.g., informal, workplace, lifelong, mobile, contextualized, and self-regulated learning). Previous editions have shown several important issues in this field, such as behavior and embodiment of pedagogic agents, suitable support of self-regulated learning, appropriate balance between learner control and expert guidance, design of personal learning environments, contextual recommendations at various levels of the learning process, tracking affective states of learners, harmonization of educational and technological standards, processing big data for learning purposes, predicting student outcomes, adaptive learning assessment, and evaluation of personalized learning solutions. PALE workshop offers an opportunity to present and discuss a wide spectrum of issues and solutions.

At this sixth edition, we were especially interested in the enhanced sensitivity towards the management of vast data coming from learners' interactions (e.g., sensor detection of affect in context) and technological deployment (including web, mobiles, tablets, tabletops), and how can this wide range of situations and features impact on modeling the learner interaction and context. Furthermore, we aimed to cover the every time more demanding need of personalized learning in wider contexts ranging from daily life activities to massive open online courses (MOOCs).

The higher-level research question addressed in this edition was: “Which approaches can be followed to cater for the increasing amount of information available from immediate (e.g., in terms of wearable devices) to broader contexts in order to provide effective and personalize assistance in learning situations?” This question has been considered in various contexts: interactive, personal, and inclusive learning environments.

PALE 2016 edition included (but was not limited to) the following topics related to personalization of learning environments:

- Affective computing
- Big data in education
- Personal and context modeling
- Data processing within and across learning situations
- Ambient intelligence
- Personalization in MOOCs
- Learning recommendation and explanations
- Learner and context awareness
- Cognitive and meta-cognitive scaffolding
- Social issues in personalized learning environments
- Open-corpus educational systems
- Adaptive mobile learning
- Reusability, interoperability, scalability
- Evaluation of adaptive learning environments
- Wearable devices for sensing and acting in ubiquitous learning scenarios
- Inclusive and adaptive education

3. CONTRIBUTIONS

A peer-reviewed process has been carried out to select the workshop papers. Three members of the Program Committee with expertise in the area have reviewed each paper. As a result, 12 submissions (out of 16) were accepted, which discuss ideas and progress on several interesting topics, such as emotional engagement, affective states, personality, deep learning and complex skills, learning analytics and recommendations, educational data mining, lifelong learning, open learner models, and adaptive MOOCs.

Arroyo et al. [2] present results of a randomized controlled study that compared different types of affective support messages delivered by pedagogical agents. Results suggest that using a character that is empathic and emphasizes the malleability of intelligence and the importance of effort provides useful results on
student learning, while reducing boredom and anxiety. Emphasizing success and failure appears to be detrimental to learning and interest and promotes anxiety.

Alyuz et al. [3] focus on the problem of emotional engagement through a personalized and multi-modal approach, and propose to detect important affective states of a learner in real time. The results show that for instructional sections, generic appearance classifier yields higher accuracy; whereas context-performance classifier is more accurate for assessment sections. Moreover, the results indicate that expression of engagement is person-specific through both of these sources, and personalized engagement models perform more accurately.

Gimenez et al. [4] deal with the use of low cost and low intrusive devices to gather contextual data to slowly drive the actions of an Intelligent Tutoring System (ITS) without constructing a fully structured model of the student and their corresponding affective and behavioral states. The idea is to improve the learning outcome and satisfaction of the student by progressively learning how to adapt the ITS in terms of the sensed data.

Huang et al. [5] explore modeling student knowledge in complex learning activities where multiple skills are required at the same time, such as in the programming domain. Their experiments show that the proposed model, based on skill combination patterns, significantly increases mastery inference accuracy and more reasonably distributes students’ efforts comparing with traditional Knowledge Tracing models and their non-hierarchical counterparts. It is a step towards building skill application context sensitive model of students’ deep, robust learning.

Okpo et al. [6] investigate what characteristics can be considered for the selection of the exercise for learners and how humans adapt exercise selection to learner personality and performance, so that an ITS can tailor exercise difficulty to these characteristics. Participants were shown an example of exercises, which would be given to learners, and asked to select the exercise which they thought the learner should do next. They responded based on the personalities of the learners as well as their past performances.

Alhathli et al. [7] investigate the influence of learner personality. In particular, it describes a study in the language learning domain that explores the relation between learners’ extroversion and the extent to which learning materials are perceived to be enjoyable and to increase their confidence and skills. They found positive correlations between extroversion and these criteria for social and active learning materials.

Vozniak et al. [8] address the problem of knowledge discovery by providing content and people recommendations based on user interests. To build a user interests profile automatically, they propose an approach by combining content analytics and activity tracking. The conducted preliminary evaluation demonstrated an ability of the approach to identify interests relevant to the user and to recommend relevant content.

Kickmeier-Rust [9] aims at developing a practical web platform that hosts tools for a theory-based approach to learning analytics. It offers tools to open and negotiate learner models using big data technologies aimed to meet the practical requirements of teachers or to really mirror human learning processes.

Alexandron et al. [10] focus on developing a general method for identifying cheaters in MOOCs in a way that does not assume a particular method of cheating. They develop a classification model that takes as input a set of features that operationalize performance and behavioral parameters that are known to be associated with cheating. These include students’ ability, the level of interaction with the course resources, solving time, and Item Response Theory person fit parameters.

Montes García et al. [11] show how the integration of a Content Management System with an adaptive framework simplifies the inclusion of personalization in existing educational applications. The use of their Within Browser Adaptation Framework (WiBAF) reduces privacy concerns, because the user model is stored on the end-user's machine. It also eliminates performance issues that currently prevent the adoption of adaptivity in MOOC platforms by having the adaptation performed within the browser.

Gilliot et al. [12] propose to explore the feasibility of personal information manager systems in the Open Learner Model (OLM) context that allows the control of personal learning data by learners themselves, its persistence and privacy. They focus on a relevant technical infrastructure giving full personal control to users without any specific competency, in order to manage long-term OLM, i.e. in lifelong and life wide perspectives.

Ishola and McCulla [13] propose an approach for supporting the lifelong professional learner. It adapts as the learner and the knowledge base change. They use data from social media to diagnose the gaps in the learner’s knowledge. The authors also try to determine what the learners know about what they know and do not know. Finally, they track how the domain of expertise is changing. The goal is to build an OLM system, wherein the gaps in the knowledge of professionals can be indicated to them at any point in time while providing personalized help also.

4. CONCLUSIONS

In this 6th edition of PALE contributions addressed several gaps identified in the state of the art, including emotional engagement, affective states, personality, deep learning and complex skills, learning analytics and recommendations, educational data mining, lifelong learning, open learner models, and adaptive MOOCs.

Nevertheless, there are many issues that remain open, such as the integration of ambient intelligence devices to gather information about the learner state while interacting in a wider range of learning settings than the classical desktop computer approach, aimed to enhance the sensitivity towards learners' interactions through diverse technological deployments (including web, mobiles, tablets, and tabletops), impacting on modeling the learner with an extended set of features (e.g. affective state) derived from their interactions and given context. We expect that future editions in PALE can progress in these directions.

5. ACKNOWLEDGMENTS

PALE chairs would like to thank the authors for their submissions and the UMAP workshop chairs for their advice and guidance during the PALE workshop organization. Moreover, we also would like to thank the following members of the Program Committee for their reviews (in alphabetical order): Hana Bydovska, Rosa Carro, Mihaela Cocea, Owen Conlan, Paul De Bra, Matt Dennis, Stefan Dietze, Sabine Graf, Tomas Horvath, Mirjana Ivanovic, Kinshuk, Noboru Matsuda, Kasia Muldner, Alexander Nussbaumer, Alexandros Paramythis, Elvira Popescu, Sergio Salmeron-Majadas, Carsten Ullrich, Peter van Rosmalen, Stephan Weibelzahl, and Naomi Wixon.

The organization of the PALE 2016 workshop has been partially supported by the following projects: Learning Layers – Scaling up Technologies for Informal Learning in SME Clusters (FP7 ICT-
REFERENCES


Addressing Affective States with Empathy and Growth Mindset

Ivon Arroyo
Worcester Polytechnic Institute
iarroyo@wpi.edu

Kasia Muldner
Carleton University
Kasia.muldner@carleton.ca

Sarah Schultz
Worcester Polytechnic Institute
schultz@wpi.edu

Winslow Burleson
New York University
wb50@nyu.edu

Naomi Wixon
Worcester Polytechnic Institute
mwixon@wpi.edu

Beverly P. Woolf
University of Massachusetts Amherst
bev@cs.umass.edu

ABSTRACT

We present results of a randomized controlled study that compared different types of affective support messages delivered by pedagogical agents. Results suggest that using a character that is empathic and emphasizes the malleability of intelligence and the importance of effort provides useful results in student learning, while reducing boredom and anxiety. Emphasizing success and failure (“That is correct/wrong”) appears to be detrimental to learning and interest and promotes anxiety. We examine a variety of student affective, cognitive and engagement outcomes in an intelligent tutoring system for mathematics.

Keywords

Empathy, Affect, Growth Mindset, Pedagogical Agents.

1. Introduction

Ideally, digital learning environments should manage the delicate balance between motivation and cognition, promoting both interest and deep learning. Students’ emotions can positively or negatively influence achievement outcomes; e.g., confidence, boredom, confusion, stress, and anxiety influence student achievement [1][2] and affective predispositions such as low self-concept and pessimism diminish academic success [3][4][5][6]. As far as science, technology, engineering, and mathematics (STEM) topics are concerned, females, minorities, and students with learning disabilities experience more frustration and anxiety when solving problems than their peers [7][8][9]. It is not surprising that these students also anticipate more barriers in STEM activities and more bias in their self-assessments [10][11]. Understanding how the environment might address negative emotions is especially important since it is experienced by most students, at various points in their learning.

Teachers attend to the affective needs of individual students [12][13], but they have limited means to recognize and respond to students’ affect in a typical classroom. Given the reality of already burdened teachers and school systems, individualized education may only be achieved through adaptive tutoring technologies that supplement traditional classroom instruction.

Interest has emerged in affect-aware technologies, given the pivotal role that affect and motivation play in the success of learning activities. The overwhelming majority of this work to date, however, has focused on modeling affect, i.e., designing computational models capable of detecting how students feel while they interact with intelligent tutoring systems (ITS) [14][15][16][17][18]. While modeling affect is a critical first step in providing adaptive support tailored to students’ affective needs, very little work exists on systematically exploring the impact of interventions on students’ performance, learning, affect, and attitudes, i.e., how an environment might respond to students’ emotions (e.g., frustration, anxiety, and boredom) as they arise.

One way to address students’ affective state is to respond in affective terms, e.g., messages that support students’ motivation to persist working on a task. However, which messages should a tutoring system, or pedagogical agent, send to students? How should pedagogical agents respond to affective states or traits of negative valence (e.g., frustration, confusion, anxiety, and lack of interest)? Should students be praised when they do well? This research focuses on how a system should address students when they are not doing well, when they make mistakes or when they show disengagement

We consider different possibilities for messages that animated affective characters deliver in a tutoring system. The testbed for our work is MathSpring1, an intelligent tutor (ITS) for mathematics that personalizes math problems, provides helps using multimedia, and effectively teaches, helping students to improve in standardized test scores [19].

Figure 1. Learning companions use gestures to offer advice and encouragement (character currently showing high interest). Students can ask for hints within practice problems. Animations, videos, and worked-out examples (shown here) add to the spoken hints about the steps in a problem.

2. BACKGROUND RESEARCH

Praising Success. The traditional way for an ITS to respond to students is to report the correctness or incorrectness of their work and to congratulate them verbally or with a ‘thumbs up’ gesture when the work is correct; when the work is incorrect, a character might move its head from side to side or not show excitement.

Training Growth Mindset. Dweck’s growth mindset theory [20][21] suggests that students who view their intelligence as an immutable internal characteristic tend to shy away from academic

1 MathSpring is freely available at http://mathspring.org
challenges; whereas students who believe that intelligence can be increased through effort and persistence tend to seek out academic challenges. Students who are praised for their effort are much more likely to view intelligence as being malleable; and their self-esteem remains stable regardless of how hard they may have to work to succeed at a task. Additionally, praise for effort encourages perseverance. In our past work, we integrated learning companions (Figure 1) into MathSpring, which were able to train attributions for “success/failure”, suggesting that effort is the cause for student success, and that mistakes are merely an indication that more effort needs to be exerted in the future to master this skill. About 20 different messages transmitted the idea that intelligence is malleable, perseverance and practice are needed to learn, that making mistakes is an essential part of learning, and failure is not due to a lack of innate ability (Table 1).

Table 1. Examples of messages spoken by the characters.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Message Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empathy</td>
<td>“Don’t you sometimes get frustrated when trying to solve math problems? I do. But guess what. Keep in mind that when you are struggling with a new idea or skill, you’re learning something, and you are becoming smarter.”</td>
</tr>
<tr>
<td>Growth Mindset</td>
<td>“Keep in mind that when you are struggling with a new idea or skill, you’re learning something, and you are becoming smarter.”</td>
</tr>
<tr>
<td>Success/Failure</td>
<td>Correct → “Excellent job!” Incorrect → “Wrong. Shall we work it out on paper?”</td>
</tr>
</tbody>
</table>

In controlled randomized studies with hundreds of students, students in general and especially certain groups of students (females and students with learning disabilities) reported increased confidence levels and decreased frustration when working with learning companions that trained growth mindset in this way, compared to not receiving learning companions. In addition, student enjoyment, self-concept, and interest were higher compared to students not given learning companions, suggesting that such affective pedagogical agents can impact students’ emotional states [7][9].

Empathic Learning Companions. To date, the Mathspring learning companions have not responded to learners’ emotional states and have acted in a counseling manner regardless of student emotion. As a result, and despite positive significant effects for the overall population of students, characters seemed to have been “harmful” to a group of students (e.g., high achieving males), who had higher affective baselines at pretest time; the characters seem to have been distracters for this group of students. Characters were more effective for lower achieving students [9] and for female students in general [7]. These results suggest that affective characters should probably be different for students who are not presently frustrated or anxious (often high achieving males). One possibility is that the behavior of the characters be adaptive to the affective state of the student. For example, the empathic characters could verbally and visually display empathy after a student has reported a negative emotion in a two-phase process: The character would: (1) be empathic to a student’s emotion saying that they are feeling that same way (e.g., “Sometimes I get frustrated when solving math problems”) and (2) resolve the situation by training failure attributions and growth mindset, (e.g., however, struggling in problems is actually a good thing, because it means that we are learning something new and becoming wiser”, see Figure 2).

D’Mello and Graesser carried out close research work on empathic characters in AutoTutor, a conversational tutor that uses 3D companions to hold dialog in natural language with the student [22]. Affective AutoTutor maps dynamic assessments of learners’ affective and cognitive states with tutor actions that address boredom, confusion, and frustration, which are sensed by monitoring conversational cues and other discourse features, gross body movements, and facial features [23]. AutoTutor responds with dialog-moves with emotional facial expression and emotionally modulated speech. For example, in response to a student’s mild boredom it states: “This stuff can be kind of dull sometimes, so I’m gonna try and help you get through it.” and in response to a student’s confusion, “Some of this material can be confusing. Just keep going and I am sure you will get it”. In comparison to a non-affective tutor, AutoTutor improved learning for low-domain knowledge learners, but was less effective at promoting learning for high-domain knowledge learners. Learning gains increased with the affective tutor whereas students’ plateaued with the non-affective tutor [24]. The affective tutor resulted in a greater positive change in perceptions than did the non-affective one, and affective response was effective during the second session of use, but not during the first session [27].

Our research questions included: a) Can we achieve similar benefits using 2D characters (HTML-based) that are less realistic than 3D characters and do not use a natural language approach?; b) Are the benefits to student learning and emotion due to empathy of the companion, i.e., what kind of benefits would a less empathic, but still highly motivational companion afford?; and c) What are the results on learning and emotion of using an empathic or less empathic companion in comparison to a control companion?

Figure 2. Learning companion empathizing to self-reported anxiety in three stages: visual acknowledgement of anxiety (left); verbal acknowledgement (middle); connector and resolution via growth mindset message (right).

that indicates only success or failure?

3. METHODS

We conducted a randomized controlled study during June 2014, with seventy-one (N=71) 7th grade students in an urban district in California.

Conditions. All conditions asked students to self-report the following emotions in a five point scale: frustration (unipolar emotion scale) and confidence/anxiety (bipolar scale)
approximately every five minutes in between math problems (MathSpring does not interrupt students while solving problems).

Students were randomly assigned to three conditions of characters delivering messages. All messages were given both in audio and written form, to guarantee the likelihood that they were exposed to the message. The conditions were: 1) Success/Failure Condition: provided traditional success/failure messages and some basic metacognitive support when students made mistakes (e.g., suggest that student asks for a hint after acknowledging the answer was incorrect); 2) Growth Mindset Condition: trained attributions for success and failure by emphasizing the importance of effort/perseverance and instilling growth mindset; 3) Empathy Condition: delivered empathic messages after the emotion is self-reported (and until the next emotion self-report five minutes later), in the following way:

- If the last emotion reported had positive valence, the character visually reflected the positive emotion with a certain probability at each problem;
- When the last emotion reported had negative valence, and with a certain probability, the character first visually reflected the negative emotion; second, it reported an empathy message such as “Sometimes these problems make me feel [frustrated],” third, a connector such as “on the other hand”, last, resolved with a growth mindset message such as “I know that putting effort into problem solving and learning from the hints will make us learn and grow our intelligence”.

Table 2. Outcome variables measured in this experiment\(^2\). The questions on the pre- and post-test were answered in a 5-point scale, going from “not at all” to “very much”.

<table>
<thead>
<tr>
<th>Math Test Performance</th>
<th>Students % score on math questions that are representative of the content covered in MathSpring.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math value</td>
<td>Measure of how important students think math is. “Compared to most of your other activities, how important is it for you to be good at math?”</td>
</tr>
<tr>
<td>Math liking</td>
<td>Measure of how much students like math. “Do you like your math class?”</td>
</tr>
<tr>
<td>Learning Orientation</td>
<td>Do students have a mastery/learning or performance orientation? “Some math classes have extra-credit projects. What kind of extra projects would you most like to do?” (1 if answered “An extra-credit project where I could learn about things that interested me.” 0 otherwise)</td>
</tr>
<tr>
<td>Learning Goal</td>
<td>Measure of how much of a learning goal students have when doing math (2 questions). “When you are doing math exercises, is your goal to learn as much as you can?”</td>
</tr>
<tr>
<td>Performance Approach Goals</td>
<td>“Do you want to show that you are better at math than your classmates?”</td>
</tr>
</tbody>
</table>

Table 3. Means and Standard Deviations of total number of messages of different kinds, seen by students in each condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Total Empathy Messages Seen</th>
<th>Total Growth Mindset Messages Seen</th>
<th>Total Success/Failure Messages Seen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empathy (N=14)</td>
<td>8.7 (3.0)</td>
<td>16.8 (7.4)</td>
<td>21.6 (12.1)</td>
</tr>
<tr>
<td>Growth Mindset (N=11)</td>
<td>0 (0)</td>
<td>20.9 (8.9)</td>
<td>28.8 (10.9)</td>
</tr>
<tr>
<td>Success/Failure (N=12)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>33.7 (14.7)</td>
</tr>
</tbody>
</table>

4. RESULTS

We analyzed the system log’s descriptive statistics to determine which messages each group had received. All students were presented with Success/Failure messages to some extent, but students in the Growth Mindset condition should have been presented with more messages regarding growth mindset training (reflections of effort and meaning of mistakes) and students in the Empathy condition should have had access to empathy messages, along with some degree of growth mindset messages and success/failure messages.

Table 3. Means and Standard Deviations of total number of messages of different kinds, seen by students in each condition.

There were some important details regarding the three conditions, namely: 1. The Success/Failure condition provided a response after getting the answer correct, and also after the second mistake made (as the first incorrect attempt triggered flashing the hint button); 2. The growth mindset condition provided one of a series of growth mindset messages after the second mistake (as the first incorrect attempt also made the hint button flash), and occasional growth mindset messages at the beginning of a new problem; however, the growth mindset condition occasionally provided some success/failure messages, as we did not want the characters to “preach” too much at every incorrect/correct attempt; for this same reason, the characters acted any response with a certain probability; 3. The empathy condition “empathized” at the beginning of a new problem with a certain probability, using both empathy and growth mindset messages as described before; however, it also used some success/failure messages as well as growth mindset messages after correct and second incorrect attempts, similar to the Growth Mindset Condition.

Procedure. Students received a pre-test on day 1, used the system on days 1, 2 and half of the 3rd day, and took a post-test starting at the middle of the 3rd day. The measures used on the pre-and post-tests are shown in Table 2. Both pretest and posttest included mathematics questions, one for each area of knowledge covered in the tutoring system. The experiment was carried out at a distance (researchers were not present at the moment of the study). Instead, teachers were instructed to run the software. Students were matched with a character of their same gender, as this had resulted in a higher learning in previous studies with middle school students, and we expected that adding a gender mis-match would add further noise to our data [7]. As measuring gender effects was not the main goal of our study, we decided to match students to a character of their same gender.

First, we measured the total number of messages seen, Table 3. Only students in the “Empathy” condition saw expressions of empathy given by characters (either visual, verbal, or both), and that students in both the “Empathy” and “growth mindset” conditions had access to (verbal) growth mindset messages. All groups saw some level of “Success/Failure” messages, with the
Success/Failure condition seeing slightly more Success/Failure messages, on average.

We next analyzed pre to posttest gains. Unfortunately, we lost pre- and post-test data because of technical issues, which left us with N=37 students with full data. We also had to match pre- and post-test data with within tutor data. Due to the low number of cases per each of the three conditions (total N=37), we decided not to carry out a cross-sectional between subjects comparison. Instead we analyzed partial correlations between exposure to the different messages indicated above and outcome variables (Table 2). We analyzed partial correlations between the total number of messages seen and a variety of post-test measures, while controlling for the corresponding pre-test measure, time spent in the tutor and total exposure to the character. The partial correlations between the total number of messages seen of each type and the post-test measures, each controlling for the corresponding pre-test measure, is shown in Table 4. We also accounted for exposure to the tutor (time spent in tutor overall) and exposure to the characters (total amount of visual or verbal character messages that the student was exposed to).

Table 4. Partial Correlations Between Specific Message Type and Post Test Measures for N=37 students, after accounting for the corresponding pre-test baseline, exposure to the tutor (time spent in tutor), and exposure to the characters (total messages heard of any kind delivered by the characters).

<table>
<thead>
<tr>
<th>Variable Measured After Using MathSpring</th>
<th>Total Empathic Messages Seen(^1)</th>
<th>Total Growth Mindset Messages Seen(^1)</th>
<th>Total Success/Failure Messages Seen(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math Post-Test Performance</td>
<td>0.34*</td>
<td>0.31+</td>
<td>-0.39*</td>
</tr>
<tr>
<td>Math Valuing Posttest</td>
<td>0.13</td>
<td>0.13</td>
<td>-0.16</td>
</tr>
<tr>
<td>Math Liking Posttest</td>
<td>0.25</td>
<td>-0.11</td>
<td>-0.14</td>
</tr>
<tr>
<td>Learning Orientation Posttest</td>
<td>0.10</td>
<td>0.26</td>
<td>-0.20</td>
</tr>
<tr>
<td>Performance-Oriented Goals</td>
<td>-0.24</td>
<td>-0.34*</td>
<td>0.33+</td>
</tr>
<tr>
<td>Frustration Posttest</td>
<td>-0.17</td>
<td>-0.09</td>
<td>0.17</td>
</tr>
<tr>
<td>Confidence Posttest</td>
<td>0.11</td>
<td>-0.10</td>
<td>-0.04</td>
</tr>
<tr>
<td>Anxiety Posttest</td>
<td>-0.43*</td>
<td>-0.16</td>
<td>0.40*</td>
</tr>
<tr>
<td>Interest Posttest</td>
<td>-0.16</td>
<td>-0.19</td>
<td>0.21</td>
</tr>
<tr>
<td>Boredom Posttest</td>
<td>-0.48**</td>
<td>-0.11</td>
<td>0.41*</td>
</tr>
<tr>
<td>Excitement Posttest</td>
<td>-0.22</td>
<td>-0.02</td>
<td>0.18</td>
</tr>
</tbody>
</table>

\(p<0.1, *p \leq 0.05, **p \leq 0.01\)

\(^1\) Only students in the empathy condition had values higher than zero for this variable
\(^2\) Students in the success/failure condition had zero values for this variable
\(^3\) Students in all conditions had values higher than zero for this variable

We can observe significant correlations between the frequency of exposure of specific kinds of messages and outcomes. A significant correlation between math posttest and empathic messages received (after partialing out for incoming knowledge as expressed in the math pretest and general exposure to the tutor and characters) indicates that students learned more as they received more empathic messages. A near significant correlation between growth mindset messages seen and math posttest performance indicates a similar trend of learning more as more growth mindset messages are seen. On the other hand, success/failure messages are negatively correlated to math posttest suggesting that the more success/failure is emphasized, the less learning will be exhibited.

A significant correlation between performance-approach orientation and growth mindset messages received indicates that more exposure to growth mindset messages is related to lower performance-oriented goals; on the other hand, high frequency of success/failure messages appears to increase performance-oriented goals. After accounting for students’ incoming math anxiety and general exposure to tutor/characters, the levels of anxiety reported by students after using the tutor were negatively correlated to exposure to empathic messages, suggesting that seeing more empathic messages would help to decrease students’ anxiety. Conversely, high frequency of success/failure messages is correlated to higher anxiety reports. Similarly, after accounting for baseline boredom towards math and exposure to the tutor and characters, boredom reported at posttest time was negatively correlated to exposure to empathic messages, and instead, boredom was positively correlated to the frequency of success/failure messages received.

5. DISCUSSION

Some of our results align to expectations in the literature: characters delivering growth mindset messages (e.g., helping students to focus on personal progress and reflect on their errors) reduced performance-approach goals (e.g., to beat classmates in comparison to a norm instead of a self-referenced focus). This suggests that growth mindset messages work according to what they were supposed to accomplish. What was not expected was that growth mindset messages would provide an apparent boost in student math learning. Similarly, we expected that empathic characters would help decrease students’ anxiety and boredom. This is consistent with results from D’Mello and Graesser [27], though for the overall tutoring session and with using simpler 2D characters instead of 3D characters. Learning companions that showed empathy helped with students’ negative affective states, in particular anxiety and boredom. Again, we did not necessarily expect that exposure to empathic messages would yield higher math performance and learning, but having characters deliver empathic messages appears to provide a boost in student math learning with the tutor. More importantly, we did not expect that success/failure messages would be so harmful to students. Regardless of whether messages indicated success or failure, the more students are exposed to these kinds of messages, the higher boredom and anxiety they develop, the higher performance-oriented goals; on the other hand, high frequency of success/failure messages is correlated to higher anxiety reports. Similarly, after accounting for baseline boredom towards math and exposure to the tutor and characters, boredom reported at posttest time was negatively correlated to exposure to empathic messages.

We observe significant correlations between the frequency of exposure of specific kinds of messages and outcomes. A significant correlation between math posttest and empathic messages received (after partialing out for incoming knowledge as expressed in the math pretest and general exposure to the tutor and characters) indicates that students learned more as they received more empathic messages. A near significant correlation between growth mindset messages seen and math posttest performance indicates a similar trend of learning more as more growth mindset messages are seen. On the other hand, success/failure messages are negatively correlated to math posttest suggesting that the more success/failure is emphasized, the less learning will be exhibited.

A significant correlation between performance-approach orientation and growth mindset messages received indicates that more exposure to growth mindset messages is related to lower performance-oriented goals; on the other hand, high frequency of success/failure messages appears to increase performance-oriented goals. After accounting for students’ incoming math anxiety and general exposure to tutor/characters, the levels of anxiety reported by students after using the tutor were negatively correlated to exposure to empathic messages, suggesting that seeing more empathic messages would help to decrease students’ anxiety. Conversely, high frequency of success/failure messages is correlated to higher anxiety reports. Similarly, after accounting for baseline boredom towards math and exposure to the tutor and characters, boredom reported at posttest time was negatively correlated to exposure to empathic messages, and instead, boredom was positively correlated to the frequency of success/failure messages received.

5. DISCUSSION

Some of our results align to expectations in the literature: characters delivering growth mindset messages (e.g., helping students to focus on personal progress and reflect on their errors) reduced performance-approach goals (e.g., to beat classmates in comparison to a norm instead of a self-referenced focus). This suggests that growth mindset messages work according to what they were supposed to accomplish. What was not expected was that growth mindset messages would provide an apparent boost in student math learning. Similarly, we expected that empathic characters would help decrease students’ anxiety and boredom. This is consistent with results from D’Mello and Graesser [27], though for the overall tutoring session and with using simpler 2D characters instead of 3D characters. Learning companions that showed empathy helped with students’ negative affective states, in particular anxiety and boredom. Again, we did not necessarily expect that exposure to empathic messages would yield higher math performance and learning, but having characters deliver empathic messages appears to provide a boost in student math learning with the tutor. More importantly, we did not expect that success/failure messages would be so harmful to students. Regardless of whether messages indicated success or failure, the more students are exposed to these kinds of messages, the higher boredom and anxiety they develop, the higher performance-oriented goals; on the other hand, high frequency of success/failure messages is correlated to higher anxiety reports. Similarly, after accounting for baseline boredom towards math and exposure to the tutor and characters, boredom reported at posttest time was negatively correlated to exposure to empathic messages.

In response to our initial research questions we have the following answers: a) we can achieve similar learning and emotional benefits with a 2D HTML-based character that is less realistic and does not process natural language; b) learning and emotional benefits are due to empathy, first, and second, some (less important) benefits are due to growth mindset messages, specifically directed at performance/learning orientation; and c) a tutor that indicates only success or failure is harmful to students, at least in comparison to other messages that emphasize the learning process and the importance of effort.

3 The data loss had to do with the pre/post-test surveys not being launched correctly by the software and incomplete student surveys in Survey Monkey, in which students failed to correctly enter their usernames or when mistakes occurred when typing.
We conclude that characters within learning environments, such as intelligent tutoring systems, are either explicitly or implicitly powerful transmitters of affective messages, with repercussions that can be shown in students’ affective states, predispositions and math learning.

6. ACKNOWLEDGMENTS

This research is supported by the National Science Foundation (NSF) #1324385 IIS/Cyberlearning DIP: Collaborative Research: Impact of Adaptive Interventions on Student Affect, Performance, and Learning. Any opinions, findings, and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

7. REFERENCES

Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario?

Nese Alyuz¹, Eda Okur¹, Ece Oktay¹, Utku Genc¹, Sinem Aslan¹, Sinem Emine Mete¹, David Stanhill¹, Bert Arnrich², Asli Arslan Esme¹

¹Intel Corporation
{nese.alyuz.civitci, eda.okur, ece.octay, utku.genc, sinem.aslan, sinem.mete, david.stanhill, asli.arslan.esme}@intel.com

²Bogazici University, Turkey
bert.arnrich@boun.edu.tr

ABSTRACT
Existing Intelligent Tutoring Systems (ITSs) are unable to track affective states of learners. In this paper, we focus on the problem of emotional engagement, and propose to detect important affective states (i.e., ‘Satisfied’, ‘Bored’, and ‘Confused’) of a learner in real time. We collected 210 hours of data from 20 students through authentic classroom pilots. The data included information from two modalities: (1) appearance which is collected from the camera, and (2) context-performance that is derived from the content platform. In this paper, data from nine students who attended the learning sessions twice a week are analyzed. We trained separate classifiers for different modalities (appearance and context-performance), and for different types of learning sections (instructional and assessment). The results show that different sources of information are generically better representatives of engagement at different sections: For instructional sections, generic appearance classifier yields higher accuracy (55.79%); whereas context-performance classifier is more accurate for assessment sections (63.41%). Moreover, the results indicate that expression of engagement is person-specific through both of these sources, and personalized engagement models perform more accurately: When person-specific data are added to the training set, on instructional sections, 85.44% and 96.13% accuracies are achieved for appearance and context-performance, respectively. For assessment sections, the accuracies are 75.25% (appearance) and 90.24% (context-performance). When only person-specific data are employed during training, similar accuracies are achieved even with very limited data.

CCS Concepts
• Human-centered computing → Human-computer interaction
• Human-centered computing → Personal computing.

Keywords
Emotional engagement, adaptive learning, personalization, affective computing, Intelligent Tutoring Systems.

1. INTRODUCTION
Current educational systems are designed based on the needs of an industrial society [1]: “one-size-fits-all”. Personalization (“accommodate-for-each”) is a key to design systems capable of addressing needs of individual students in the Information Age [2]. Technology is considered as an enabler for personalization in education [3]. Towards this end, Intelligent Tutoring Systems (ITSs) are used to track the learning process of students by monitoring their actions, creating a learning profile for each student, and providing real-time feedback for many learning difficulties [4], [5]. Although such systems are capable of personalization to some extent, they lack the required empathic capabilities. We envision a novel technology - an empathic autonomous ‘tutor’ - playing a role similar to a 1:1 human tutor.

Relevant research indicates that engagement is positively correlated with learning [6]: The more students are engaged in learning activities, the more they learn. In [7], the overall engagement level of a student is defined as a combination of three parameters: (1) Cognitive engagement, defining the inner psychological quality during the learning process; (2) behavioral engagement, representing the learner’s observable actions (e.g., OnTask/OffTask); and (3) emotional engagement, corresponding to affective states of a learner during a learning task (e.g., happiness, boredom, or confusion). In this research, we see emotional engagement detection as a trigger for personalized experience. The majority of current ITSs provide teachers with a rough idea of students’ engagement in learning tasks based on interaction data between student and the content platform. However, they do not provide any concrete information about students’ affective (i.e., emotional) states – especially throughout instructional tasks.

In our research, our goal is to develop the empathic autonomous ‘tutor’ that can closely monitor students in real-time using multiple sources of data to understand their affective states. We aim to use affective state information as a trigger for personalizing students’ learning experience: For example, if a student is reading an article and identified as bored, a video representation of the content can be suggested. Or, if the student is detected as confused while solving a math question, hints can be provided for scaffolding.

The remainder of this paper is organized as follows: In Section 2, the related literature is reviewed, considering the most important challenges driving our work. The proposed methodology is outlined in Section 3, followed by a summary of our experimental results in Section 4. Section 5 highlights our conclusions and future directions for our research.

2. RELATED WORK
Although there are some efforts towards affective computing in education, some major challenges still remain unaddressed. Such challenges include the following: (1) Learning-related affective states should be considered instead of the six basic emotions [8]; (2) Data acquisition in real-life scenarios is a challenging task [9]; (3) Multimodal approaches should be employed for improved engagement modeling [10]; and (4) Model personalization is...
necessary for accurate detection [11]. The remainder of this section will describe each of these challenges with literature.

2.1 Learning-related affective states

Students’ affective states can influence overall learning outcomes either positively or negatively [12], [13], [14]. Recognizing and addressing such states is crucial to positively impact student learning [15], [16]. However, in a classroom where there is one teacher and many students, addressing those states for every individual in a timely manner is often unrealistic. This brings up the need for intelligent systems capable of detecting and taking actions towards students’ affective states.

There is an extensive track of research on detecting facial expressions specifically focusing on basic emotions of anger, fear, sadness, happiness, disgust, and surprise as described by Ekman [17] (an exhaustive review can be found in [18]). However, a recent review of 24 studies shows that the six basic emotions are not directly applicable to learning domain [12]. Instead, affective states such as bored, confused, satisfied (i.e., delight) are commonly observed during learning [19].

There are studies focusing on the development of intelligent systems that can automatically detect students’ affective states and intervene accordingly to induce positive learning outcomes [12]. For example, in [20], the binary classification problem of whether a student was interested or not during learning activity was investigated. In [21], [22], and [23], automatic recognition of frustration was investigated. In [24], students’ posture is used to track boredom (low engagement) and flow (high engagement). In [25], affective states considered are confused, frustrated, engaged, bored, or neutral. In their most recent study [9], an updated list of affective states was divided into three categories: positive (i.e., ‘Excited’, ‘Calm’), neutral (i.e., ‘Satisfied’), and negative (i.e., ‘Bored’, ‘Confused’). The contextual features are explained in detail in Section 3.2. However, in literature, there are only a limited number of studies including contextual information as a source of modality. In [21], the multimodal sensory information from facial expression was combined with information about a learner’s activity on a computer. In [10], the use of physiological data were employed together with the contextual data from the tutoring system.

In our research, we employ contextual information coming from the content platform as the complementary modality to appearance. Here, the platform provides us information about the context (e.g., difficulty level of the content) and the performance (e.g., number of hints taken). The contextual features are explained in detail in Section 3.2.

2.2 Data acquisition in real-life scenarios

Another challenge we aim to address in this research is data acquisition in real-life scenarios. Although there has been a great interest in detecting learning-related affective states, most of these studies are limited in terms of learning scenarios and/or amount of data used. In the majority of such studies, the data collection took place in a controlled laboratory environment. The advantage of lab environments is the controlled ambience (e.g., lighting, or background) with minimal distractions [9]. However, data collected in such environments does not allow to create models that capture real complexities of classrooms. In the literature, there are only a few studies that employ in-the-wild (i.e., in a realistic classroom scenario) databases [10], [9]. Unfortunately, these databases are limited in terms of data (i.e., 4-5 one-hour long sessions), and none are publicly available. Therefore, for our research, we collected and labeled approximately 210 hours of student data collected over 17 sessions in authentic classroom scenarios.

2.3 Multimodal approach

Environmental factors in authentic classrooms usually result in far noisier data compared to data collected in a lab, especially for appearance data. Due to distractions in real-life classroom scenarios, appearance information (i.e., face-related data) can be polluted with unusual head pose or hand gestures obstructing facial area, and thereby preventing detection based on facial features. Moreover, according to [27], identical facial configurations include significantly different emotions depending on context. It is also indicated in [28] that the interpretation of human behavioral signal requires one to know the context where it is displayed, since its interpretation is significantly context-dependent. For instance, satisfaction can be expressed as a smiling face in context of leisure time and as a neutral face with wider-eyes in context of learning, whereas a frowning action unit can be an indication of anger in context of inter-personal communication and frustrated in context of learning. Therefore contextual information from a content platform can act as a complementary source of information. However, in literature, there are only a limited number of studies that employ contextual information as a source of modality. In [21], the multimodal sensory information from facial expression was combined with information about a learner’s activity on a computer. In [10], the use of physiological data were employed together with the contextual data from the tutoring system.

In our research, we employ contextual information coming from the content platform as the complementary modality to appearance. Here, the platform provides us information about the context (e.g., difficulty level of the content) and the performance (e.g., number of hints taken). The contextual features are explained in detail in Section 3.2.

![Figure 1. Overall scheme of the generic emotional engagement detector.](image-url)
2.4 Model personalization

Despite of individual differences, current efforts in affect-related educational research are towards creating a generic learner-state model. However, as shown in recent research [11], [29], [30] for basic emotion recognition, personalized models can perform better in terms of detection accuracy. To address this gap in the educational research, we propose to develop a personalized emotional engagement model. Details on the proposed personalization scheme are provided in Section 3.

3. METHODOLOGY

In this paper, we aim to develop a system that can detect a learner’s emotional engagement through a personalized and multi-modal approach. The system setup includes a student using a computing device (such as a PC or a tablet) equipped with a camera, and consuming educational content through a content platform. The overall scheme for the generic emotional engagement detector is given in Figure 1: The raw data acquired from camera and content platform are fed into corresponding feature extractors, and then to the classifiers. The two classification outputs define the emotional engagement status of the learner.

Once the affective state of a student is detected in an online manner, emotional engagement information can be used either directly by providing interventions for an improved learning experience of the student or by showing the teacher real-time status for each individual student through a dashboard.

3.1 Data Collection and Labeling

3.1.1 Data Collection

One of the major challenges is to collect labeled data that is necessary for model training. Towards this end, we ran authentic classroom pilots with real students from a high school in Turkey. Our target group was 9th grade students (14-15 years old). We collected data in a math course that was optionally offered as a part of this study to interested students. The lessons were scheduled twice a week with 17 sessions for 20 students in total. Among the 20 volunteering students, three of them dropped the course within the semester. Overall, nine of the students participated in the course twice a week, whereas the other eight participated once a week. At the end, around 210 hours of data were collected from these students.

Students used an online, publicly available math learning tool as a content platform in the sessions. Each of the sessions took around 60 minutes. During these sessions, the students watched instructional videos related to different math topics in the school curriculum and solved exercises (i.e., math questions) related to the topics covered. For each session, a math teacher was present in the classroom as a mentor. The specific curriculum was selected by the teacher as being appropriate for student level.

During each session, our data collection framework recorded the video of the individual students with a 3D camera (i.e., Intel® RealSense™ Camera F200), and collected the context and performance logs from the content platform. Each student worked independently in the class using a laptop computer.

3.1.2 Labeling Process

For the supervised training phase of our models and for the performance evaluation of our system, ground truth labels were necessary. The labeling was done by experts with a background in educational psychology. We incorporated Human Expert Labeling Process (HELP) as described in [31] to rigorously label student data with respect to affective states. We developed and utilized a labeling tool. The experts provide labels based on inspecting four different inputs: They simultaneously monitor individual students’ videos and corresponding desktop captures; listen to audio including environment noise and students’ voices; and view additional contextual information about the recording (e.g., session number, lecture topic, etc.) to decide on final labels. The experts did continuous labeling: Whenever they observe a state change in student’s data, they assigned a new label.

To perform labeling, eight labelers were hired and trained by an educational researcher using HELP. For increased reliability, each recording of a student was labeled by five different labelers. The inter-rater reliability was measured after the training session and was regularly tracked during the labeling process to detect any outliers.

For the emotional engagement, we initially followed the suggestion from [8] and used four labels, each corresponding to one quadrant of the circumplex model – ‘Excited’, ‘Calm’, ‘Bored’ and ‘Confused’. In addition to these four affective states, we also used the label ‘Unknown’ stating that the labeler cannot decide on the state, and the label ‘N/A’ (Not Available) stating that a segment is not valid either due to student is not visible or class-content is not active.

The results from the labelers’ post-interviews showed that a distinction between positive and negative arousal for positive valence states was not clear. In [8] it was proposed that from an educational point of view, the two positive valence quadrants can be treated in the same way. Following this suggestion, we merged the positive valence states of ‘Excited’ and ‘Calm’ into a single state ‘Satisfied’. To reinforce this decision, the inter-rater agreement between the original label set with one label per quadrant and the adapted with a single label for positive valence quadrants were compared. As the reliability coefficient for measuring the agreement among multiple raters, we utilized Krippendorff’s alpha [32], and the results are given in Section 4.1.

3.1.3 Final Label Assignments

After the recorded data were labeled by the five labelers, we analyzed the labelers’ decisions. Note that we used sliding windows of 8-seconds (with an overlap of 4-seconds) and treated each window as a separate instance. Hence, the labeling data with intervals defined by each labeler for each state-change, was divided into fixed instances with duration of 8-seconds. To assign a final label to each instance, majority voting is applied together with validity filtering (i.e., if there is no majority among labelers, an instance is labeled as invalid/unknown). In addition to the final labels, an agreement level for each instance is assigned to carry out further experiments with data belonging to different agreement levels. Numerical details are provided in Section 4.1.

3.2 Feature Extraction

The features used in our system refer to the segments of 8-seconds length with an overlap of 4-seconds.

3.2.1 Appearance Features

The videos of individual students were recorded with Intel® RealSense™ Camera F200 during the data collection sessions. The raw video data includes the RGB and depth streams which are used for the extraction of low-level features via Intel® RealSense™ SDK [33]: Face location and head pose in the 3D space, 2D and 3D positions of 78 facial landmarks, head pose, 22 facial expressions, and seven basic facial emotions together with sentiment values are
considered as the frame-wise features. These frame-wise features are only employed in the extraction of segment-wise higher-level features necessary for emotional engagement detection. As in [34], we extracted higher-level features including various L-estimator statistical values (e.g., tri-mean of head velocity) and energy calculations (e.g., trend of pose energy), related to head position and pose, to facial expressions, and to seven basic emotions. These robust statistical features constitute the appearance features. The groupings of appearance features used in this paper are given in Table 1.

Table 1. Appearance and context-performance feature subgroups and the corresponding feature counts.

<table>
<thead>
<tr>
<th>Appearance Features</th>
<th>Number of Features</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracking ratio</td>
<td>2</td>
<td>Position and pose tracking</td>
</tr>
<tr>
<td>Head position and pose</td>
<td>128</td>
<td>Trend of pose energy, median of absolute head center acceleration, standard deviation of head position, etc.</td>
</tr>
<tr>
<td>Facial expressions</td>
<td>32</td>
<td>Number of right eye raisers per segment, mean of smile, etc.</td>
</tr>
<tr>
<td>Seven basic emotions</td>
<td>28</td>
<td>Mean of anger intensity, number of joyful segments etc.</td>
</tr>
<tr>
<td>TOTAL</td>
<td>190</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context-Performance Features</th>
<th>Number of Features</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time related</td>
<td>6</td>
<td>Time from beginning, video/attempt duration, etc.</td>
</tr>
<tr>
<td>Trial related</td>
<td>3</td>
<td>Trial number, number of trials until success, etc.</td>
</tr>
<tr>
<td>Hint related</td>
<td>5</td>
<td>Number of hints used on attempt or question, etc.</td>
</tr>
<tr>
<td>Grade related</td>
<td>7</td>
<td>Grade, correct attempt percentage, etc.</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>Gender, question number from beginning, etc.</td>
</tr>
<tr>
<td>TOTAL</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

3.2.2 Context and Performance Features

Contextual features are extracted partly from user profiles and session information we had in our database (i.e., gender, age, time of a day), in addition to the data from the content platform (i.e., video duration, exercise/trial number, time within a session). Some of these contextual features are related to the instructional sections (e.g., video duration), some are related to the assessment sections (e.g., question number), and some are related to both types of these sections (e.g., time from beginning). The performance features are extracted from the user profile data containing user characteristics provided by the content platform. Note that these performance features are all related to the assessment sections, in which the students are expected to solve exercises. These features are extracted either per each assessment section where a group of questions are solved in a row, per each question or per each attempt (i.e., each trial within a question). In general, performance features are related to the grade, the time spent, the number of trials and the number of hints taken for a question. In addition to these initial performance features, we examined features that were used in [35] with a different content platform. We adapted those features that are applicable to our platform.

Since contextual and performance features are both obtained from the content platform, we employed data fusion at feature level and concatenated these two features into one context-performance feature set. The groupings of context-performance features examined in this study are given in Table 1, together with feature counts and some exemplary features.

3.3 Uni-modal Classification

As the uni-modal classifiers, we employed Random Forest (RF) classification method [36]. The idea behind RF is that it grows many decision trees while using a randomly selected subset of training data for each tree. Moreover, a randomly selected subset of features is used to split each node. The final class for a test sample is assigned by the majority vote among all trees. The advantages of using RF as a classifier are that there is no need for pruning and cross validation, and over-fitting is not an issue. For all these reasons, Random Forest (RF) with 100 classification trees is selected to be the final classification method. For the two different modalities, we trained two separate RF classifiers: (1) Appearance classifier, and (2) context-performance classifier.

3.4 Model Personalization

As empirically shown in [29], [30], person-specific models achieve significant improvement over person-independent models if the subject-specific data are sufficient for model training. For our future work, we envision to obtain personalized models for emotional engagement detection through online self-labeling of the person-specific data: During the lecture, the individual students will be asked through intervention pop-ups to self-label themselves at randomized time points. The self-labeling interface will also be embedded into the content platform and will be permanently reachable, so that the students can give labels any time. At the end of each session, the training data will be updated with the labeled person-specific data and the model will be retrained.

In this paper, we investigated the improvement that could be achieved by personalized models: Since self-labels are not available for the current dataset, we considered the ground truth labels as self-labels. We applied personalization through including person-specific data during the training phase, for both of the uni-modal classifiers. We experimented with two different approaches on how to include the person-specific data: (1) ‘Adapted’, and (2) ‘Personal’. In ‘Adapted’, we augmented the initial training set of the ‘Generic’ model, collected from a different set of students, with the acquired and labeled instances of the test subject. In ‘Personal’, person-specific training sets are generated by using only the personal data. The aim of the ‘Adapted’ model is to merge the capabilities of the ‘Generic’ detector (which is trained on a large database) with the characteristics residing in the person-specific data. However, if the personal data is sufficient for training, the ‘Personal’ model would be better to represent person-specific behaviors. In Figure 2, the training sets used in different models are visualized. In Section 4.3, the preliminary experiments to show the need for personalized models are summarized.

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Training sets used for different models: (1) Generic, (2) Adapted, and (3) Personal.
4. EXPERIMENTAL RESULTS

For emotional engagement detection, we experimented with two different modalities: (1) Appearance and (2) context-performance. The learning content included two different types of sections: (1) Instructional, and (2) assessment. Currently, these two section types are considered as two separate problems, and two separate models are constructed. For each different modality and each section type, we experimented with three different models: (1) ‘Generic’, where the training set contains data of students separate from the test subject in a leave-one-out manner; (2) ‘Adapted’, where the training set is augmented with the subject-specific data; and (3) ‘Personal’, where subject-specific training sets are constructed using only the subject-specific data.

4.1 Labeler Agreement Analysis

As outlined in Section 3.1.2, during the labeling phase, the four emotional labels, ‘Excited’, ‘Calm’, ‘Bored’, ‘Confused’, and two other labels ‘Unknown’, ‘N/A’ were used (i.e., 4+2 states). As stated in the labelers’ post-interviews, the distinction between high and low arousal for positive valence states was not always clear. This experience was in line with the proposal of [8] to treat positive valence quadrants in the same way. Following this suggestion, we merged the two positive valence states ‘Excited’ and ‘Calm’ into a single state ‘Satisfied’ (i.e., 3+2 states). To reinforce our decision, we compared the inter-rater agreement between the original 4+2 label set and the adapted 3+2 one: ten hours of data collected from four subjects were labeled by three persons in both ways. The inter-rater agreement level according to Krippendorf’s alpha [32] was 0.2 in the original 4+2 label set and it increased to 0.4 in the adapted 3+2 label set.

Since multiple labelers were employed for the labeling process, it was necessary to assign a final label to each of the instances. For this process, we applied majority voting. However, since emotional labeling is a subjective task and the inter-rater agreement level for the small experimentation outlined above is not sufficiently high (below 0.8), we conducted a filtering over the traditional majority voting: We computed the ratio of the agreement, and grouped instances accordingly. We had three agreement levels: (1) High, (2) medium, and (3) low; for 5/5, 4/5, and 3/5 majority votes, respectively. The other samples were regarded as instances of disagreement. The data distribution of agreement levels are visualized in Figure 3, in (a) for instructional, and in (b) for assessment samples. As can be seen in these figures, the number of ‘Confused’ samples for the instructional, and the number of ‘Bored’ samples for the assessment sections were too few. Therefore, we have discarded ‘Confused’ class for the instruction sections, and ‘Bored’ for the assessment sections. Furthermore, we decided to use High-Medium agreement level for the instructional sections. However, for the assessment, we used all of the agreement levels, since the assessment sections were short in general and this would have led to limited number of samples in total. In addition to agreement samples, we included the disagreement instances as representatives of the ‘Unknown’ class.

To investigate how the performance of the personalized models changed, we selected students who attended most of the sessions (i.e., twice a week). Therefore, in the experiments summarized in this paper, the data from nine of the students are utilized.

4.2 Generic Classification Results

Although the main problem addressed in this paper is model personalization, we also included results on the generic model for comparative purposes. For each section type (instructional vs. assessment) and for each modality (appearance vs. context-performance), separate RF classifiers are trained. The available data of each student are divided into training and test sample sets, as approximately 80% and 20% of the whole data, respectively. For each individual, we carried out leave-one-subject-out approach, where the training samples of all the other students are utilized to construct the training set of that individual’s classifiers. Due to data imbalance, we also experimented with 10-fold random down-sampling to construct balanced training sets: For each student, the instance count of the limiting class (with the minimum number of training samples) is used for random instance selection per class, and the random selection is carried out for ten times. As results, we reported F1 measure which incorporates both precision and recall values.

4.3 Personalization Results

In this paper, we employed the ground truth labels to construct person-specific labeled sets, necessary for the personalization experiments. When constructing person-specific models, we considered two approaches: (1) ‘Adapted’, where the training set of the generic model is augmented with the person-specific data; (2) ‘Personal’, where only the person-specific data is used in the training phase. The results for the three models of ‘Generic’, ‘Adapted’, and ‘Personal’ are compared in Table 2 and Table 3, for instructional and assessment sections, respectively. The average numbers of test instances for each student are given in column 2. The average number of instances used in training, and average F1 values for the appearance and the context-performance classifiers are given in columns 3-5, in columns 6-8, and in columns 9-11; for ‘Generic’, ‘Adapted’, and ‘Personal’ models, respectively. As the
5. CONCLUSIONS AND FUTURE WORK

The aim of this work is to detect emotional engagement of a student while the learner is consuming educational content. In this paper, we investigated appearance and context-performance modalities. For a better understanding of the classification performance, we treated instructional and assessment sections separately. For the different modalities and different section types, we experimented with three models: (1) ‘Generic’, (2) ‘Adapted’, and (3) ‘Personal’. The results of the generic models showed us that appearance is more informative for the instructional sections (55.79% vs. 49.50%), whereas with the presence of performance-related features for the assessment sections, context-performance modality becomes more representative (63.41% vs. 48.12%). As the personalization experiments indicated, information included in both of the modalities are person-specific, thus model personalization is a must to obtain highly performing emotional engagement models. Context-performance classifiers achieve high improvement even with limited personal data (90.89-97.32%), whereas improvement for the appearance modality is lower (76.37-89.30%) and requires more person-specific instances to achieve accuracies as high as context-performance. For both modalities, ‘Generic’ models can be used for emotional engagement detection, if no person-specific data are available. Through acquisition of personal data, however, ‘Adapted’ models should be utilized. After sufficient amount of person-specific data are collected, ‘Personal’ models should be preferred for improved accuracies.

In future work, we will investigate fusion strategies to merge appearance and context-performance modalities. To further increase the dataset volume and to validate personalization experiments with real self-labels, we are currently designing a new data collection pilot. We are redesigning course content so that data imbalance can be decreased while increasing samples for ‘Confused’ and ‘Bored’ classes. Moreover, we are investigating strategies for a better representation of the ‘Unknown’ class. We are planning on including bio-sensors as an additional modality.

We are working on strategies for personalization, where the need for manual or self-labeling is minimized. Moreover, we are working on generic and personalized feature selection methods to identify important traits.

6. REFERENCES


[34] L. Nanxiang and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in Int. Conf. on Multimedia and Expo (ICME), 2013.


Affective and Behavioral Assessment for Adaptive Intelligent Tutoring Systems

L. Marco-Gimenez, M. Arevalillo-Herraez, F.J. Ferri, S. Moreno-Picot
Dept. Informatica. Universitat de Valencia. Spain
margi4@alumni.uv.es,{miguel.arevalillo,francesc.ferri,salvador.moreno}@uv.es

J.G. Boticario, O.C. Santos, S. Salmeron-Majadas, M. Saneiro, R. Uria-Rivas
aDeNu, Dept. of AI, Comp. Sci. School. UNED. Madrid. Spain
{jgb,ocsantos,sergio.salmeron,marsanerio,raul.uria}@dia.uned.es

David Arnau
Dept. of Didactics of Math. Universitat de Valencia. Spain
david.arnau@uv.es

José A. González-Calero
Dept. of Mathematics. Univ. of Castilla-La Mancha. Spain
jose.gonzalezcalero@uclm.es

Raúl Cabestrero, Pilar Quirós
Faculty of Psychology, UNED. Spain
{rcabestrero,pquiros}@psi.uned.es

Aladdin Ayesh
Faculty of Technology De Montfort University, UK
aayesh@dmu.ac.uk

P. Arnau-González, N. Ramzan
School of Eng. and Computing Univ. of the West of Scotland, UK
pablo.arnau@uws.ac.uk

ABSTRACT

Adaptive Intelligent Tutoring Systems (ITS) aim at helping students going through the resolution of a given problem in a principled way according to the desired outcomes, the intrinsic capabilities of the student, and the particular context in which the exercise takes place. These systems should be capable of acting according to mistakes, boredom, distractions, etc. Several works propose different models to represent the problem being solved, the student solving it and the tutor guidance to the desired solution. The system complexity requires non trivial models whose corresponding parameters need to be estimated with different kinds of data, usually requiring heavy and difficult sensing and recognition tasks. In this work, we present some of the work in progress in the BIG-AFF project. Between other issues, we deal with the use of low intrusive devices to gather contextual data to loosely drive the actions of an ITS without constructing a fully structured student model with corresponding affective and behavioral states. The idea is to improve the students’ learning outcome and satisfaction by progressively learning how to adapt the ITS in terms of the sensed data.

Keywords

Personalization; Adaptation; Intelligent Tutoring Systems; Word Problem Solving; Affective states; Emotion detection

1. INTRODUCTION

Previous research has provided solid evidence that emotions strongly affect motivation and hence play an important role in learning [1]. In the research project BIG-AFF, we build on the hypothesis that it is possible to provide learners with a personalised support that enriches their learning process and experience by using low intrusive (and low cost) devices to capture affective multimodal data that include cognitive, behavioural and physiological information.

Results from previous research [9, 40] led to the identification of open issues that need to be addressed to advance the scientific and technological knowledge regarding emotion recognition:

- Information sources. For the potential applicability of the methods in realistic learning environments, it is essential that devices are both low intrusive and inexpensive. This leads to considering quantified-self approaches and wearable technologies to gather diverse physical and physiological information about the user. In addition, computer vision methods [14, 8, 18, 38, 21] that can be run on inexpensive and ubiquitous hardware can complement this information.

- Context Modelling. Although commonly ignored or not given the importance that it deserves, the usage context has a relevant effect on the interaction (e.g. navigation patterns depend on the input devices, keyboard, mouse, touch screen, as well as the functional diversity of the user) [17].

- Individual influence. There is a personality and physiological influence of the individual when feeling and expressing emotions. This implies the need for designing appropriate experimental settings using both
inter-subject and intra-subject configurations to better understand the user impact on the detection process.

- **Data processing.** A major limitation for real time processing is the large amount of data simultaneously produced by independent devices. Deep learning methods for multimodal data fusion [23, 35, 7, 11] may help discovering more suitable abstract data representations. In addition, Big Data approaches provide a powerful framework for dealing with flows of large sets of unstructured data. They both may open up new possibilities in the way data is mined to find emotional and behavioural patterns.

Apart from evidencing the high complexity of the problem at hand, initial findings have revealed the need to further investigate experimental methodologies and related infrastructural support to help clarify the fragile and elusive nature of affect. According to the literature [27] and our background, which is grounded on affect-detection experiences at large (nearly 100 subjects) and small scale (intra-subject), there are no yet clear criteria for setting the appropriate affect detection and management experimental approaches. Within the context of the BIG-AFF project, we are deeply investigating into what types of experiences should be carried out as well as, how and when, depending on circumstances. We are also concerned with the application of user-centric evaluation to measure the impact of affect in the users' experience and learning gains, and test ecological validity. To this end, different contexts (individual, collaborative, ambient intelligence, enriched multimedia) are being considered. As a result, we have started to develop an extensive knowledge about the organization of the experiences, reliable (expert and user) data labelling and non-intrusive techniques to automatically recognize emotions.

### 2. PREVIOUS WORK BY THE CONSORTIUM

Project participants have extensive experience at learning technologies and computer vision. On the methodological side, TORMES methodology [28] allows to elicit educational oriented recommendations; and the Collaborative Logical Framework [29] allows to create effective scenarios that support learning through interaction, exploration, discussion, and collaborative knowledge construction. On the development side, extensive work has been performed in intelligent tutoring systems (ITS) and computer vision methods. In particular, an ITS for the arithmetic and algebra domains has been developed [5, 3, 4, 13]. Moreover, the potential applicability of traditional holistic techniques in the context of facial expression recognition using a method based on eigenfaces [21] has been analysed; and works have been developed on modelling the relation between facial action units and changes in the affective state, using Kinect devices [6].

The consortium has an extensive research experience on some of the open issues identified above. This experience comes from the joint research carried out under the project MAMIPÉC [30]. This project aimed to provide affective personalized support to learners on educational contexts, trying to identify learners' affective states from a multimodal approach based on mining data gathered from several input data sources. To this end, large-scale experiments were carried out in laboratory settings (nearly 100 participants, including visual impaired people). These aimed at building a database of emotional data, from where to analyse the viability of inferring learning emotions in an educational context [31]. Multimodal approaches were also used to combine these interactions with physiological signals [32, 24]. These research experiences aimed to detect changes in the user emotional states while solving mathematical problems. They served to identify an ad-hoc methodology to tag facial expressions and body movements that conform to changes in the affective states of learners while dealing with cognitive tasks in a real time learning process [26]. They also served to explore the viability of mouse and keyboard indicators for emotions detection [25]; and of other low cost devices and specific techniques related to the field of computer vision [21]. In a later experience, affective recommendations identified with TORMES methodology [33] were validated. In addition, AICARP open platform has been implemented at low cost with Arduino to detect changes in physiological signals that can be associated with stressful situations, and when this happens, it recommends the learner to relax by delivering modulated sensorial support in terms of light, sound, or vibration at a relaxation breath rate [34].

### 3. EMOTION DETECTION

#### 3.1 Physiological sensing: EEG

In the task of identifying the emotion, numerous authors have used diverse kinds of physiological signals [15, 39, 36]. Electroencephalography (EEG) stands out as one of the most studied and yet less understood. In general, these EEG-based models are used to build a classification scheme that use EEG features as an input and yield a prediction related to the user’s emotional state as an output [2, 20]. One major factor that affects the system’s performance is related to the existing implicit relations between the selected features and the user’s reaction to changes in the variables considered in the emotional model. In this sense, feature selection has a remarkable influence on the system’s performance. As a first contribution aligned with the project objectives, we have developed a novel classification scheme that combines connectivity and energy EEG features. [36, 19, 17] The method is based on a feature reduction scheme that integrates a one-way ANOVA with a Principal Component Analysis (PCA) to yield two dimensional data which is fed into a non-linear Support Vector Machine (SVM). The method has been tested by using DEAP [16], a database commonly used as a benchmark in this type of applications. Results outperform the ones reported in [16], both in detecting valence and activation.

#### 3.2 Video sensing

Processing images and videos using low-cost devices lead to effective ways of sensing affective states [23]. But instead of accurate, general-purpose recognisers we are interested in methods able to give appropriate hints about some affective variables in real time. Gabor Transform (GT) is a signal processing tool specially suitable when recognition is invariant to motion in certain dimensions. The use of GT as an appropriate feature extractor on large video streams is also motivated by new linear time algorithms able to be used in live scenarios [22, 37, 12].
The use of GT is specially useful in the context of sensing small local changes that are correlated with some affective variables as it is the case of slight changes in a moving face in front of a camera. Videos can therefore be segmented depending on local activity, even helping to segment the video in moving objects against a fixed background, or detect movement as opening of mouth, eyes, and parts of the face, helping to detect gestures directly.

Efficient GT has also been applied to recognize the 3D structure of moving objects, i.e., a face, by taking into account that there are fixed parts (forefront, ears, hair, etc) which behave together as rigid solid subject to perspective changes. Then, it is possible to recover the lost information in a way quite similar to a Markov filter, as only 6 variables are needed to locate the head in space. In this way it is possible to present the face in a normalized 3D way which can greatly help to decide which changes are due to perspective, and which ones are truthfully due to face expressions.

4. REACTING TO EMOTIONS

One major worry in ITS is related to the most adequate level of help that should be provided to the subject to optimize learning [10]. Arguments supporting or against intensive scaffolding methods, that always allows the user to reach the end of the problem, have been common in the literature. Researchers against this type of strategy claim negative effects because students could solve problems without engaging in their content. We reported the effect of an intensive scaffolding in the learning of algebraic word problem solving in our ITS [13]. In particular, we carried out an empirical study in which the effectiveness of two versions of the ITS (one with intensive scaffolding and another without it) were compared. The results show a significantly increase of the competence in algebraic word problem solving in the group that used the ITS with intensive scaffolding in comparison with the group that used the reduced version.

Nevertheless, we have recently extended this study to observe the impact of intensive scaffolding on variables other than learning, by using an adapted version of the ITS where the user inputs his/her state after solving each problem. While the effect on valence and activation does not show a significant correlation with the level of help supplied, the autonomy shows a correlation above 0.35, that we consider significant in statistical terms (as more help, as less gain in autonomy). This means that, although intensive scaffolding may not have a negative impact in learning in a direct way, it may have it in an indirect way by acting on affective variables.

Results derived from this research lead to the necessity to design appropriate methods that simultaneously improve learning and affective variables (valence, activation and autonomy). To this end, we are currently working on incorporating classification methods that decide on the most appropriate help level for each problem step, taking into account variables related to the current student knowledge and the difficulty of the problem at hand, such as the user’s previous skill in solving problems, the number of help requests per step and the number of steps with help in the last solved problem, and information about changes in valence, activation and autonomy after finishing the current problem reported by the user through a SAM test. This information, in conjunction with precomputed average scores about users’ skill, number of help requests per step and number of steps with help for the previous and current problems, constitutes the input for a set of SVM classifiers whose aim is to predict the most adequate help level for the next problem in terms of valence, activation and autonomy. The SVM classifiers with RBF kernel were implemented in Python using scikit-learn free software machine learning library1, trained with a leave-one-out cross-validation technique, and optimized with an exhaustive Grid Search procedure to estimate the optimum C and G parameters. The parameters for each classifier and the accuracy in terms of recall and ROC area are shown in the table 1.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>C</th>
<th>G</th>
<th>Accuracy</th>
<th>ROC</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>100</td>
<td>0.01</td>
<td>70%</td>
<td>0.747</td>
<td>171</td>
</tr>
<tr>
<td>Activation</td>
<td>10</td>
<td>0.01</td>
<td>62%</td>
<td>0.643</td>
<td>179</td>
</tr>
<tr>
<td>Autonomy</td>
<td>10</td>
<td>0.01</td>
<td>64%</td>
<td>0.683</td>
<td>234</td>
</tr>
</tbody>
</table>

Table 1: Estimated parameters obtained by a Grid Search strategy, accuracy and ROC area for the implemented classifiers

At the time of writing this paper, experiments are being carried out at a secondary school to test the actual performance of these classification methods when used in a real environment. To perform the experiments, we have used a Ubuntu Linux live distro with a Xfce desktop environment with data persistence, and with the appropriate and necessary tools to run the ITS in combination with the implemented classifiers.

5. EMOTION MODELING: COMBINING GLOBAL AND INDIVIDUAL MODELS

The performance of global affective models to detect emotions is limited by the subject’s individualities [6], which are not taken into account in this type of settings. However, individual models generally suffer from the small sample problem, as there is an intrinsic difficulty associated with gathering extensive data from the same user.

We are currently working on approaches that are able to use the whole set of data, while at the same time allow for user individualities to be taken into consideration. To this end, we have used two types of methods:

- Clustering methods. As a first approach, users were grouped by common and relevant characteristics that the models are compatible, such as age or gender. However, the results obtained did not significantly improve the state-of-the-art based on real modes. Our current line of work includes methods where the clustering is learned, rather than given according to user features.

- Refinement methods. In this approach, we construct a global model, which is then processed and adapted to each individual as data becomes available. It is in this approach where we get considerable improvement with respect to both, using individual models and global models.

1http://scikit-learn.org/stable/
6. CONCLUDING REMARKS AND ONGOING WORK
Teaching arithmetic word problem solving is a complex task. The ability of a teacher to provide feedback that is consistent with the current student reasoning and affective states is a major factor to arrive at the desired learning outcomes. In this paper, we have described strategies to incorporate this kind of skills into an adaptive ITS by considering low cost, low intrusive devices.

Acknowledgments.
This work has been partly supported by the Spanish Ministry of Economy and Competitiveness through project BIG-AFF with grants TIN2014-59641-C1-2-P

7. REFERENCES
[22] S. Moreno-Picot, M. Arevalillo-Herráez, and W. D. Villanueva. A linear cost algorithm to compute the discrete Gabor transform. IEEE Transactions on...


Modeling Skill Combination Patterns for Deeper Knowledge Tracing

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

Julio D. Guerra-Hollstein
School of Information
University of Pittsburgh
Pittsburgh, PA, USA
jdg60@pitt.edu

Peter Brusilovsky
School of Information
Sciences & Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA, USA
peterb@pitt.edu

ABSTRACT

This paper explores the problem of modeling student knowledge in complex learning activities where multiple skills are required at the same time, such as in the programming domain. In such cases, it is not clear how the evidence of student performance translates to individual skills. As a result, traditional approaches to knowledge modeling, such as Knowledge Tracing (KT), which traces students’ knowledge of each decomposed individual skill, might fall short. We argue that skill combinations might carry extra specific knowledge, and mastery should be asserted only when a student can fluently apply skills in combination with other skills in different contexts. We propose a data-driven framework to model skill combination patterns for tracing students’ deeper knowledge. We automatically identify significant skill combinations from data and construct a conjunctive knowledge model with a hierarchical skill representation based on a Bayesian Network. We also propose a novel evaluation framework primarily focuses on the knowledge inference quality, since we argue that traditional prediction metrics no longer suffice to differentiate between shallow and deep knowledge modeling. Our experiments on datasets collected from two programming learning systems show that proposed model significantly increases mastery inference accuracy and tends to more reasonably distribute students’ efforts comparing with traditional KT models and its non-hierarchical counterparts. Our work serves as a first step towards building skill application context sensitive model for modeling students’ deep, robust learning.

Keywords
complex skill, multiple skill, composition effect, Knowledge Tracing, Bayesian Network, robust learning, deep learning

1. INTRODUCTION

Knowledge Tracing (KT) [6] 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 and map student problem-solving performance into student knowledge level for each of the skills. Knowledge Tracing demonstrated its ability to track student knowledge for different domains and could be now considered as the most popular learner modeling approach. However, a known limitation of Knowledge Tracing is the assumption of skill independence in problems that involve multiple (complex) skills. Recent research, however, challenged this assumption demonstrating that there is additional knowledge related to specific skill combinations, in other words, the knowledge about a set of skills is more than the “sum” of the knowledge of individual skills [13], some skill must be integrated (or connected) with other skills to produce behavior [20]. For example, students were found 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) [13].

This points out that at least in some domains, we need to pay specific attention to modeling student knowledge considering skill combinations. One of these domains is arguably computer programming. Research on computer science education and pedagogy has long argued that knowledge of a programming language can’t be reduced to a sum of knowledge about different programming constructs since there are many stable combinations or patterns (also known as schemas, plans) that have to be taught and practiced [8, 27]. To generalize to other domains, we argue that skill combination patterns can also potentially represent “chunks”, production rules, or general problem solving patterns that are critical for defining expertise in a domain.

In this paper we use datasets from SQL and Java programming tutors to demonstrate the feasibility and the value of modeling skill combination patterns. We present a data-driven framework for modeling skill combinations including a novel student model and new evaluation metrics.

2. BACKGROUND

2.1 Patterns in Programming Expertise

Experts in the area of psychology of programming have long argued that programming plans and other kinds of patterns form an important part of programming expertise [8]. Most actively, all kinds of programming patterns: plans, techniques, templates, and “cliches” were used by researchers in the area of intelligent tutoring systems (ITS) for programming to support intelligent analysis of student programs [18, 31]. While intelligent debuggers recognized and diagnosed pattern errors, they do not maintain a model of student
knowledge on pattern level. First student models on the level of patterns were introduced by Brusilovsky [1] who used expert-suggested construct pairs as knowledge components for problem sequencing and Weber [30] who applied larger programming “episodes” as knowledge components for adaptive recommendation of program examples [31]. The more advanced episodic model has never been expanded or ported to another languages due to its complexity and high knowledge engineering demands. In contrast, the simpler pair-based approach has been used in a few follow-up projects [23]. The work presented in this paper continues this stream of research focusing on lower-level skill combinations and their automatic (rather than manual) discovery.

2.2 Complex Skill Student Modeling and its Evaluation

Complex skill student knowledge modeling has been a challenge and attracts increasing attention in student modeling community. Starting from Gong [9] constructing variants based on traditional Knowledge Tracing using multiplication or minimization among skills, more advanced models have been put forward to address the multiple skill credit and blame assignment issue [22, 32, 11, 33], and many of them resemble cognitive diagnosis models such as DINA or NIDA [19]. However, these student models only use a “flat” knowledge structure which might overlook the actual important dependency or interaction among skills. Among the works that consider hierarchies or relations among skills, most of them focus on prerequisite relations [5, 2] or granularity hierarchy [25, 12], which differ from skill combination relation in nature and cannot be readily applied. Also, most of these works rely heavily on expert laboring, and automatic methods to discover skill relations is still a recent endeavor [4].

Regarding the data-driven evaluations of complex skill student models, prior works mostly uses prediction performance [32, 11, 33]. There is a growing concern of only using prediction performance for evaluating student models. [10] have shown that highly predictive model may be useless for adaptive tutoring, and they can have low parameter plausibility or consistency [17]. While some attempts have been made to evaluate in terms of the effect for tutoring [22], a recent learner outcome-effort paradigm [10] offers a promising way to empirically evaluate student models for adaptive tutoring. In this work, we extend both of our previous frameworks [10, 17] for evaluating complex skill student models.

3. PROPOSED FRAMEWORK

This section introduces our framework of incorporating skill combination patterns for deeper knowledge tracing. It consists of model construction and model evaluation.

3.1 Model Construction

We construct a Bayesian Network (BN), Conjointive Knowledge Modeling with Hierarchical Skill Combination (CKM-HSC), to model skill combinations. Figure 1 shows the structure of our model. It supports three functionalities: 1) performance prediction, 2) knowledge estimation, and 3) mastery decision. The O nodes (shaded) represent observed binary student performance, K nodes represent binary latent skill knowledge level, and M nodes represent the aggregated binary latent skill knowledge level that we call Mastery. Particularly, we introduce Mastery nodes, in order to reflect the idea of granting skill mastery for each skill based on all the skill combinations’ knowledge levels. Edges denote causal relation among the variables. One major challenge of applying BN for complex skill modeling is the time and space complexity. We made simplifications in the model to balance complexity and accuracy, which we further explain.

Figure 1: Bayesian Network Structure of Conjunctive Knowledge Modeling with Hierarchical Skill Combinations

3.1.1 Skill Combination Pattern Representation

We represent skill combination patterns in a hierarchical way in the Knowledge and Mastery parts (see Figure 1):

I The first layer consists of basic individual skills. It captures the basic understanding and application of a skill. For example, in Java, we can have K1 representing the basic understanding and application for ForStatement, and similarly, K2 for ArrayElement.

II The intermediate layers consist of skill combinations which can be derived from smaller skill units. These layers capture a deeper understanding of each individual skill considering when and how to apply it with other skills in more complex situations. For example, K1,2 can represent the joint skill of ForStatement and ArrayElement requiring iterating through an array. As a first step to simplify the problem, we only consider skill combinations from two basic individual skills.

III The last layer represents the Mastery in each of the individual skills, where the nodes are fed from skill combinations or single skills. We introduce this layer in order to reflect the idea of granting skill mastery for each skill based on all the skill combinations’ knowledge levels. To avoid repeated computation, combined skills only connect to one Mastery node, the one representing the latter basic skill in the temporal order in which the skills appear in the course.

Knowing skills from lower layers serves as prerequisites for knowing skills in higher layers. We argue that such a dependency (hierarchy) is crucial. For example, for an item requiring skill combinations, if a student fails, the model can differentiate whether the problem is in the basic skill or due to a lack of experience of applying skills together; if a student succeeds, the model may increase the belief that the student already knows the basic individual skill through the prerequisite link.

To model the relation between basic individual skills and combined skills, we use multinomial distributions, since different basic individual skills might have different importance in affecting the knowledge of combined skills. However, this can impose exponential complexity. In this framework we limit skill combinations to be composed from an upper bound of N lower level skill units, typically choosing N as a
small number. In the case where high order combinations are involved, we can consider using causal independence models to reduce to linear complexity.

3.1.2 Learning Network Structure and Parameters

Since the network involves latent variables, we use Expectation Maximization algorithm to conduct network learning. The final structure of the network depends on which skill combinations are incorporated. If we don’t limit the search space of skill combinations, it will scale exponentially. So we employ some heuristics to select skill combinations. Algorithm 1 outlines a greedy search algorithm. It requires a pre-ordering of the skill combination candidates. During each iteration, it compares the cost functions (e.g., data log likelihood) of the network with a skill combination incorporated with the optimal one from previous iterations. However, in each iteration, the posteriors of latent variables have to be computed which can be very time-consuming.

Algorithm 1 Learn CKM-HSC

Input: potential skill combination list C, original item to individual skill adjacency matrix Q (qmatrix), initial skill to skill adjacency matrix H (all zero), student performance data matrix O, initial parameters Θ
Output: new Bayesian network B with final structure, fitted parameters, and posteriors of latent skills

1: B = ConstructBN(Q, H, Θ)
2: B, Cost = LearnBN(B, O)
3: C’ = SelectAndRankSkillCombinations(C, O)
4: for each skill combination pattern c ∈ C’ do
5: Q’ = UpdateQmatrix(Q, c)
6: H’ = UpdateHmatrix(H, c)
7: B’, Cost’ = LearnBN(B’, O)
8: if Cost’ < Cost then
9: B = B’
10: Cost = Cost’
11: end if
12: end for
13: return B

To increase the run time efficiency, we construct a simplified version replacing the search procedual with empirical thresholding (pruning), as shown in Algorithm 2 which we use in the current work. Preliminary analysis with a subset of the data (Java database with 158 students on 45 items) showed that we can achieve comparable results with the simplified version of the algorithm and reduce computation time significantly (from 9 hour to 30 minutes to finish structure and parameter learning) based on the popular Bayes Net Toolbox [24]. We leave for the future to explore faster implementation tools and alternative techniques (e.g., approximate inference) to address this issue.

Now we introduce how we select skill combinations and thus learn the BN structure in Algorithm 2. Firstly, we get all the possible pairwise combinations of skills by each pair’s co-occurrence in each item, which results in a big list of skill combination patterns. Then, we apply following criteria (with higher importance criterion ordered before) through the algorithm to select the skill combinations (note that in order to get ranked skill combinations for an item (Line 14: function GetRankedSkillCombinationsForItem), we also use criterion I-II).

Algorithm 2 Learn CKM-HSC (simplified)

Input: C, Q, H, O, Θ (same as in Algorithm 1); item to possible skill combination matrix P, thresholds α1,4
Output: new Bayesian Network B with final structure, fitted parameters, and posteriors of latent skills

1: C’ = [ ]
2: for each skill combination pattern c ∈ C do
3: βc = EstimateDifficulty(c, P, O)
4: S = GetIndividualSkills(c)
5: βS = EstimateDifficulties(S, Q, O)
6: if (βc - Max(βS)) > α1 and βc > α2 then
7: C’ = Add(C’, c)
8: end if
9: end for
10: M = GetItemList(Q)
11: for each item m ∈ M do
12: βm = EstimateDifficulty(m, O)
13: if βm > α3 then
14: Cm = GetRankedSkillCombinationsForItem(m, P, O)
15: C’m = [ ]
16: for each combined skill c ∈ Cm do
17: if c ∈ C’ and Length(C’m) ≤ α4 then
18: C’m = Add(C’m, c)
19: else
20: break
21: end if
22: end for
23: Q’ = UpdateQmatrix(Q, m, C’m); Q = Q’
24: H’ = UpdateHmatrix(H, C’m); H = H’
25: end if
26: end for
27: B’ = ConstructBN(Q’, H’, Θ)
28: B’ = LearnBN(B’, O)
29: B = B’
30: return B

I The difference of difficulties between the combined skill and its hardest individual skill should be large and non-zero (Line 6: βc - Max(βS)). Otherwise, the original individual skills should be able to capture the difficulty of an item already and extra skill is not needed.

II The difficulty of the combined skill should be high (Line 6: βc > α2).

III An item with higher difficulty is more needed to be refined, in this case, to be indexed with combined skills (Line 13: βm > α3).

IV Each item is indexed with a limited number of skill combinations (Line 17: Length(C’m) ≤ α4). Our preliminary study shows that the number of skill combination candidates increases quickly even with a small relaxation of this criterion.

In order to estimate difficulty of skills or items (Algorithm 2 the function EstimateDifficulty or EstimateDifficulties), one possible way is to apply some existed models such as Item Response Theory [29] or Additive Factors Model [3] to extract related parameters. However, any such pre-estimations can’t avoid imposing assumptions of learning or skill relations. For example, IRT assumes no learning while our datasets are collected from learning activities; AFM assumes compensatory relation among skills while on our datasets a conjunctive relation is more suitable. We use a simple heuristic method which also imposes assumptions.
but we argue it should be no worse than the above alternatives given its simplicity of computation. We estimate item $m$’s difficulty $\beta_m$ by computing the average error rate across students on this item. For estimating skill $s$’s difficulty $\beta_s$, we apply a $Min$ gate (get the minimum value) to the difficulty estimations ($\beta_m$) of all items ($M$) containing this skill, assuming each skill’s basic difficulty level is determined by the easiest item in which the skill is required, i.e., $\beta_s = \min_{s \in m, 1 \leq m \leq M} (1 - \beta_m)$. We use this formula for both estimating an individual skill’s difficulty and a skill combination’s difficulty.

Learning BN with hierarchical structures among latent skills based on temporal learning data is non-trivial. To simplify the modeling process, we ignore the temporal learning effect during the model learning process, while maintaining the dynamic knowledge estimation power during the application phase (see 3.1.3). Such simplifications have been used in many prior works [5, 28], and we argue that it can be reasonably compensated by the dynamic updating process (see subsection 3.1.3). We leave for future work to include the temporal learning effect in our model.

3.1.3 Dynamic Knowledge Update

At each practice opportunity, the learned network provides the inference of a latent knowledge node. For each student’s first practice, the network uses the same priors for latent knowledge nodes to perform inference, and only when we update the knowledge by different student’s performance, the network starts to differentiate among students by maintaining different up-to-date knowledge estimates. In order to achieve this, CKM-HSC follows the same dynamic BN roll-up mechanism in [5]: it uses posterior knowledge probabilities conditioned on historical observations as the priors for next time step. Note that we only need to update the basic individual skills in the first layer. The states of other skill variables will be computed based on fitted conditional probabilities. The following formulas outline the update of the knowledge state of a basic individual skill $s$ at time step $t$ after observing an evidence:

$$P(K_{i,t+1}=\text{known})_{\text{prior}} = P(K_{i,t}=\text{known})_{\text{posterior}}$$ (1)

$$= P(K_{i,t}=\text{known})O=\text{evidence}$$ (2)

This updating procedure differs from KT [6]: it ignores the transition probabilities between time steps. However, similar to [5], we argue that the change in knowledge estimates is mainly determined by the new evidence. We leave for the future incorporating learning parameters.

3.1.4 Performance Prediction

CKM-HSC applies a Noisy-AND gate for each item’s conditional probability distribution, assuming that students need to know all the underlying skills in order to succeed given our datasets’ nature on which only one answer is accepted. Noisy-AND gate and other causal independence models has been used in many prior works [5], particularly in the popular psychometric model DINA [19]. The major benefit is that it only requires two parameters (guess and slip) for each item, regardless of the number of skills required by the item.

This significantly reduces the complexity of the model. Note that for items having skill combinations, the probability of getting this item correct only depends on the combined skills and is independent of the basic individual skills. Equations 3 and 4 show the CPT of item $i$ consisting of its guess and slip probabilities $g_i$ and $s_i$:

$$P(O=\text{correct}|\text{all skills=known}) = 1 - s_i$$ (3)

$$P(O=\text{correct}|\text{at least one skills=unknown}) = g_i$$ (4)

3.1.5 Mastery Decision

To access the mastery level of each skill, CKM-HSC aggregates knowledge estimates from each skill combination assigned to current skill (if no skill combinations are present, then the basic skill), and gives a final knowledge level of a skill, based on which, skill mastery is decided. It means that to reach mastery, students need to know both, skill’s basic meaning and how to correctly apply it with other skills. We aggregate by computing the joint probability of all required skill units being in known state as the probability of a mastery node as shown in Equation 5. Since skill combinations share parents, computing this joint probability considers the dependencies among skills:

$$P(M_i=\text{known})=P(K_{i,1}=\text{known}... K_{i,j}=\text{known})$$ (5)

where $K_{i,1}$ to $K_{i,j}$ denotes the skill combinations assigned to $M_i$. If a skill has no assigned skill combination, then

$$P(M_i = \text{known}) = P(K_i = \text{known}).$$ (6)

3.2 Model Evaluation

We argue that modeling “deeper” knowledge of complex skills requires also “deeper” evaluation, and only examining prediction performance is not enough, as explained in Section 2.2. We propose a new evaluation framework extending a recent Learner Effort-Outcome Paradigm (LEOPARD) [10], and a multifaceted evaluation framework [17]. We think that in a real world tutoring system relying on a student model’s knowledge inference to provide support, knowledge inference quality should be of primary importance, and also parameter plausibility shouldn’t be overlooked.

Mastery Accuracy (knowledge inference quality). The basic idea is that once a student model asserts a student’s mastery for an item’s required skills, the student should be very unlikely to fail. In the original metric, each single skill’s knowledge state is examined and accuracy is computed on data after reaching mastery threshold. This is not applicable to multiple skill case since the responsibility of each skill for the actual performance is not clear. To address this, we examine the multiple skill knowledge states jointly as shown in Algorithm 3.

Mastery Effort (knowledge inference quality). This metric empirically quantifies the number of practices students needed to reach a level of mastery of the set of skills in a domain inferred by a student model. It is computed based on each individual skill’s expected effort as shown in Algorithm 4 (which is similar to the original metric in [10]). To apply LEOPARD to multiple skill models, we make a simplification: each practice count can be treated as a count for each individual skill involved. We argue that it won’t change the relative effort comparing among models because all models will be computed under the same simplification. In a preliminary study on our data set, we find out that only
one out of 20 students has reached mastery for all attempted skills. This suggests that computing effort based on per skill perspective is better since it allows to have sufficient data. We leave for the future to further improve this metric.

Algorithm 3 Get mastery accuracy (multiple skill)

Input: vector $Y$ of actual correctness performance (ordered by time step within each student), matrix $K$ of a student model’s inferred knowledge levels (probability of known) for required skills per observation (with the same order as $Y$), item to skill matrix $Q$, mastery threshold $p$

Output: mastery accuracy metric

```plaintext
1: NbObsWithSkillsInferredMastery = 0
2: NbObsCorrectAndSkillsInferredMastery = 0
3: for each observation $y_i \in Y$ do
4:     $s_t = \text{GetRequiredAggregatedIndividualSkills}(y_i, Q)$
5:     $k_i = \text{GetInferredKnowledge}(t, s_t, K)$
6:     AllSkillsInferredMastery = TRUE
7:     for each skill’s inferred knowledge $k_{q,t} \in k_i$ do
8:         \begin{align*}
9:             & \text{Judge whether current required skill is inferred mastery: if any one skill is not inferred mastery, change the flag and stop checking:} \\
10:             & \text{if } k_{q,t} < p \text{ then} \\
11:             & \text{AllSkillsInferredMastery = FALSE; break}
12:             & \text{end if}
13:             & \text{end for}
14:             & \text{end if}
15:             & \text{end for}
16:             & \text{if NbObsWithSkillsInferredMastery > 0 then}
17:             & \text{MasteryAccuracy = } \frac{\text{NbObsCorrectAndSkillsInferredMastery}}{\text{NbObsWithSkillsInferredMastery}}
18:             & \text{return } \text{“No sufficient data.”}
19:             & \text{end if}
20:             & \text{return MasteryAccuracy}
```

Two important considerations for computing the metrics described above are explained as follows.

- **The balance between Mastery Accuracy and Effort.** As our later experiments shown, it seems that a model can achieve higher mastery accuracy by letting students practice for longer time. However, we argue that having an acceptable level of Mastery Accuracy (e.g., > 0.85) is necessary for a real world tutoring system, and thus it is needed and worthy that more practice effort is required before reaching this level of accuracy. Otherwise, the system will risk granting mastery in few practices when students haven’t reached it, which we think is more problematic than delaying the granting of mastery.

- **Mastery Threshold.** Different mastery thresholds have been used in prior works [6, 10]. Typically 0.95 is used, yet the rationality behind it is also not clear. We present evaluations on an extensive range of mastery thresholds [0.5, 0.99], but primarily focus on higher threshold regions (e.g., ≥ 0.7).

- **Amount of data to compute mastery metrics.** As shown in Algorithm 3 and 4, to compute mastery accuracy or effort, we need to have data with corresponding knowledge states inferred as mastery state by student models. When mastery is not reached (by inference), LEO-ARD uses imputation, which we think is not suitable on thresholds with few students inferred to reach mastery. This might distort the original distribution. So, we remove imputation, and only focus on thresholds with sufficient data with at least 25% of the complete data available to compute the mastery metrics.

Algorithm 4 Get mastery effort (multiple skill)

Input: $K$, $Q$, $p$ (same as in Algorithm 3)

Output: mastery effort metric

```plaintext
1: MasteryEffort = 0
2: for each skill $q$ in the complete skill list do
3:     MasteryEffortPerStu$q$ = ]
4:     for each student $u$ attempted items requiring $q$ do
5:         MasteryEffort$u,q$ = 0
6:     Select inferred skill $q$’s knowledge level sequence corresponding to student $u$’s response sequence:
7:         $k_{u,q} = \text{Filter}(K)$
8:     for each practice $t \in [0, \text{Length}(k_{u,q})]$ do
9:         \begin{align*}
10:             & \text{Judge whether the current practice a student is inferred reaching mastery of a skill:} \\
11:             & \text{if } k_{u,q,t} < p \text{ then} \\
12:             & \text{MasteryEffort$u,q$ += 1} \\
13:             & \text{else} \\
14:             & \text{Otherwise, stop counting:} \\
15:             & \text{break}
16:             & \text{end if}
17:             & \text{MasteryEffortPerStu$q$ = Average(MasteryEffortPerStu$q$)}
18:             & \text{Add(MasteryEffortPerStu$q$, MasteryEffort$u,q$)}
19:             & \text{end for}
20:             & \text{MasteryEffort += MasteryEffort$u,q$}
21:             & \text{end for}
22:             & \text{MasteryEffort = Average(MasteryEffort$u,q$)}
23:             & \text{Accumulating number of practices to get the total number of practices to reach mastery of all skills for a student on average:} \\
24:             & \text{MasteryEffort += MasteryEffort$u,q$}
25:             & \text{end for}
26:             & \text{return MasteryEffort}
```

**IDI (parameter plausibility).** To examine parameter plausibility, we utilize the Item Discriminative Index (IDI) used in psychometric models for evaluating item qualities and refining q-matrix [7]:

$$IDI_i = 1 - g_i - s_i$$  \(7\)

where $g_i$ and $s_i$ denotes item $i$’s guess and slip probabilities. A higher IDI is preferable which have both low guessing and slipping rates. We utilize this metric to evaluate models with item level guess and slip probabilities.

**RMSE, AUC (performance prediction accuracy).** We report two popular prediction metrics used in evaluating student models, Root mean squared error (RMSE) and Area Under the Receiver Operating Characteristic curve (AUC) based on a recent paper [26]’s suggestion.
Table 1: Dataset descriptive statistics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#obs.</th>
<th>#items</th>
<th>#skills</th>
<th>avg #skills/item</th>
<th>#users</th>
<th>#correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL</td>
<td>17,197</td>
<td>45</td>
<td>34</td>
<td>0.70 (.01)</td>
<td>366</td>
<td>58%</td>
</tr>
<tr>
<td>Java</td>
<td>25,988</td>
<td>45</td>
<td>56</td>
<td>0.72 (.01)</td>
<td>347</td>
<td>67%</td>
</tr>
</tbody>
</table>

4. STUDIES

In this section we describe studies that demonstrate CKM-HSC’s advantage over traditional and alternative models, give a closer examination of extracted skill combinations, and briefly explore the effect of adding external knowledge.

4.1 Datasets and Experimental Setup

We used two datasets, SQL and Java programming, collected through classroom studies between Fall 2013 and Fall 2015 at the University of Pittsburgh. The SQL dataset is from an online system SQL-KNOT, and the Java dataset is from the system JavaGuide [15]. Both systems contain problems requiring students to apply multiple skills at the same time, and they allow students to have multiple attempts where each attempt corresponds to a new instantiation of the same template. In SQL-KNOT, students are asked to write complete SQL statements to solve problems, and only one solution is accepted; in JavaGuide, students are requested to predict output of different Java programs and only one answer is accepted. Both systems assess correctness (0/1).

Problems are grouped by topics and indexed by a set of fine-grained concepts by experts assisted with ontologies and an automatic Java parser [14]. For computational efficiency, on Java dataset, we remove two most complex topics (Interface and Inheritance), and randomly selected 45 items. Table 1 shows the descriptive statistics of the final datasets used (with multiple attempts).

One both datasets we conducted a 10-fold student stratified cross-validation. In each fold we trained on 90% of students, and predicted the probability of getting a problem correct and inferred the probability of knowing a skill for each observation of the remaining 10% of new students. On each fold, we evaluate each model from following multi-grained concepts by experts assisted with ontologies and an automatic Java parser [14]. For computational efficiency, on Java dataset, we remove two most complex topics (Interface and Inheritance), and randomly selected 45 items. Table 1 shows the descriptive statistics of the final datasets used (with multiple attempts).

On each fold, we evaluated each model from following multi-grained metrics introduced in Section 3.2:

- **Knowledge inference quality**: Mastery Accuracy, Mastery Effort.
- **Parameter plausibility**: IDI
- **Performance prediction accuracy**: RMSE, AUC

For each metric, we compute the average value across 10 folds and 95% confidence interval based on t-distribution. Note that during predicting on test set, we perform dynamic update: each model gradually gather information of the positive range among different application context, risking students with combinations or not, and blindly distributes students’ effort differently updates the skill whether an item requires skill combinations, WKT in-duced on thresholds with enough data points to compute mastery accuracy. We set convergence criterion as 10

β = 0.3. We set α as the 75th percentile’s value of the positive range of the difference (β - Max(β3)) computed on all skill combinations based on the performance data (ordered from small to large), and similarly, we set α2 as the 75th percentile of β2, and βm values computed from the data. We set α2 = 2. These values of α are set based on our preliminary studies consulting experts. On average (across 10 folds), our automatic method extracts 14 and 30 skill combinations on SQL and Java datasets.

4.2 Is proposed skill combination incorporated model better than traditional KT models?

In the first study, we compare following models:

- **KT-Single**: Each item is mapped to one coarse-grained skill (topic). This is an implementation of the classic Knowledge Tracing [6] which uses a Hidden Markov Model for each skill to model the latent knowledge and predict the performance of students. The hidden variable represents knowledge level (learned or unlearned) and the observable represents performance (correct or incorrect).
- **WKT**: Each item is mapped to a set of skills (concepts). We fit each skill independently as classic Knowledge Tracing Model and then take the minimum of each skill’s predicted probability of success as the final prediction. We only update the knowledge of this weakest skill when the evidence is an incorrect response, while we update all skills when the evidence is a correct response. This is a model used in many prior works [9].
- **CKM**: Our proposed conjunctive knowledge modeling without incorporating skill combinations. It fits, predicts and updates skills jointly by Bayesian Network.
- **CKM-HSC**: Our proposed model. Each item is mapped to a set of skills, which can be either individual skill or skill combination with hierarchy among them. It fits, predicts and updates skills jointly using a BN.

On both datasets, CKM-HSC has comparable predictive performance as other models (Table 2). However, CKM-HSC has significantly better mastery accuracy than other models (on thresholds with enough data points to compute mastery metrics and high enough values to be considered as proper mastery thresholds), and it also estimates that more effort is needed to reach mastery (Figure 2). As mentioned in Section 3.2, we think that having such more practices in CKM-HSC is necessary in order to reach an acceptable mastery accuracy.

We are still interested in comparing how CKM-HSC and WKT distribute students’ effort, so we further conduct a drill-down effort analysis on SQL dataset (Figure 3). For skills that potentially have skill combinations, WKT indifferently updates the skill whether an item requires skill combinations or not, and blindly distributes students’ effort among different application context, risking students with shallow learning reaching mastery, and also it directs students to spend more effort on skills without combinations. CKM-HSC clearly distributes students’ efforts as follows: it
4.3 Is using hierarchy better than independence for incorporating skill combinations?

In this study, we investigate the effect of using hierarchy for incorporating skill combinations. We compare following three models:

- **WKT-SC**: Using WKT’s framework, adding skill combinations as new independent individual skills.
- **CKM-SC**: Using proposed model’s framework, adding skill combinations as new independent individual skills.
- **CKM-HSC**: Our proposed model, adding skill combinations in a hierarchical way.

Figure 4: CKM-HSC vs. alternatives to incorporate skill combinations on SQL. Grey lines denote regions with enough data points to compute mastery metrics and high enough values to be considered as proper mastery thresholds.

We mainly report results on SQL dataset. Three models have similar predictive performance (RMSE=0.47±0.01, and AUC=0.7±0.01). However, comparing other metrics, we see important difference among the models (Figure 4).

Comparing WKT-SC with CKM-HSC, the latter has better mastery accuracy and similar effort on higher mastery thresholds.

A drill-down analysis (Figure 4) shows that WKT-SC tends to require students to spend much more effort on basic individual skills' understanding for skills with combinations and also tends to require more mastery effort on skills without combinations; while CKM-HSC tends to require students to put major effort in skill combinations. Comparing CKM-HSC with CKM-SC, we can see that including hierarchy significantly improves mastery accuracy and requires more mastery effort, and slightly improves parameter plausibility from an IDI value of 0.34±0.01 to a value of 0.35±0.01. Similarly, we observe that the different behaviors among three models are not simply the effect of setting different mastery thresholds.

On Java dataset, we reach similar conclusion: all models have similar predictive performance, but CKM-HSC improves mastery accuracy and requires similar mastery effort, and it also has higher IDI than CKM-SC.
Table 3: Comparison among CKM-HSC and alternatives adding external knowledge on Java (average across 10 folds with 95% CI; mastery metrics computed on [0.7, 0.93]).

<table>
<thead>
<tr>
<th>Models</th>
<th>M.Acc</th>
<th>M.Effort</th>
<th>IDI</th>
<th>RMSE</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKM-HSC</td>
<td>.84</td>
<td>83</td>
<td>.46(0.01)</td>
<td>.45(0.01)</td>
<td>.73(0.01)</td>
</tr>
<tr>
<td>CKM-HSC-P</td>
<td>.82</td>
<td>70</td>
<td>.45(0.01)</td>
<td>.45(0.01)</td>
<td>.73(0.01)</td>
</tr>
<tr>
<td>CKM-HSC-P-E</td>
<td>.82</td>
<td>59</td>
<td>.42(0.01)</td>
<td>.45(0.01)</td>
<td>.73(0.01)</td>
</tr>
</tbody>
</table>

4.4 Can we improve modeling by adding external knowledge for skill combination extraction?

Our impression from manual inspection of the automatically extracted skill combinations from data was mostly positive. For example, on Java dataset, many extracted skill combinations contain Loop related concepts (such as ForStatement) combined with IfStatement, AddAssignment-Expression, or Array related concepts. However, we also observed extracted pairs which are not apparently meaningful. For example, java.lang.String.substring combined with java.lang.System.out.print. We conduct further analysis to investigate whether using external information to increase skill combinations’ interpretability can improve student modeling. We compare three models on the Java dataset:

- **CKM-HSC**: Our proposed model only using performance data to extract skill combinations. This results in 41 skill combinations from 10 folds.
- **CKM-HSC-Proximity**: Extracted pairs are further constrained by proximity, i.e., we only consider pairs appearing in the same line or in the same block (for example a skill inside a for-loop and the for-loop are in the same block). This results in 31 skill combinations.
- **CKM-HSC-Proximity-Expert**: Experts mark pedagogically meaningful pairs from the skill pairs extracted by CKM-HSC-Proximity. This results in 19 skill combinations.

We summarize the results in Table 3. Interestingly, adding extra knowledge sacrifices mastery accuracy slightly, and also decrease models’ parameter plausibility, but it saves much more effort. This shows that purely data-driven approach, despite some human interpretability issues, may capture some latent, implicit relationship among skills to enable it has the highest mastery accuracy and parameter plausibility, but it may require much more mastery effort. A promising direction is to find a balance between data-driven approach and human interpretability.

5. CONCLUSIONS

In this paper, we examined the importance of skill combinations in domains of multiple-skill problems where multiple skills can be integrated (combined) together to require extra knowledge or produce extra difficulty not captured by contributing individual skills. Our work is the first attempt to model this skill combination effect using data-driven techniques. We constructed a Conjunctive Knowledge Model with Hierarchical Skill Combinations (CKM-HSC) based on Bayesian Networks. In our new model, mastery of a skill can only be granted when a student demonstrates ability to apply this skill with other skills in varied contexts. We demonstrated that incorporating skill combinations can significantly increase mastery assertion accuracy and more reasonably direct students’ practice effort comparing with traditional knowledge tracing models and its non hierarchical counterpart. We propose an evaluation framework for student models built for multiple skill knowledge modeling. Under our new evaluation framework, we effectively quantify the accuracy and expected mastery effort of a student model’s knowledge inference, which can not be addressed by traditional performance prediction metrics.

We also addressed the problem of computational complexity by using suitable network representation and heuristic data-driven methods. In the future, we will explore more efficient implementation tool and new techniques.

We are aware that it is very challenging to learn the structure of a complex hierarchical, multi-layer Bayesian Network, so we will consider collecting larger datasets, and datasets with more sparse connections among nodes. We will also consider consulting experts for determining (refining) the structures of the network, and giving structure priors of the network.

Our current study only considers skill combination in pairs, and it will be to interesting to consider more complex skill combinations, and even consider difficulty factors from cognitive task analysis [21]. In the longer term, we expect that such a framework can be extended to model skill application context, “chunks”, schemas or patterns for defining expertise or deep learning of a domain [16].

We expect that our new skill combination model can provide more significant benefits when deployed in real-world tutoring systems. It can potentially enable more powerful visualization for open student models and better remediation. It can encourage students to practice in more contexts on the way to mastery, and guide content authors in developing content that addresses skill combinations. Not only serving as an attempt to bridge data mining with pedagogical and learning theory, our work also raises attention in the community to build student models for deeper, robust student learning.

Acknowledgement

This research is supported by the Advanced Distributed Learning Initiative (http://www.adlnet.gov/) and the National Science Foundation Cyber-Human Systems (CHS) Program under Grant IIS-1525186.

6. REFERENCES


Exploring Requirements for an Adaptive Exercise Selection System

Juliet Okpo
University of Aberdeen
Aberdeen, UK
r02jao15@abdn.ac.uk

Matt Dennis
University of Aberdeen
Aberdeen, UK
m.dennis@abdn.ac.uk

Judith Masthoff
University of Aberdeen
Aberdeen, UK
j.masthoff@abdn.ac.uk

Kirsten A. Smith
University of Aberdeen
Aberdeen, UK
r01kas12@abdn.ac.uk

Nigel Beacham
University of Aberdeen
Aberdeen, UK
n.beacham@abdn.ac.uk

ABSTRACT

In this exploratory focus group study, we investigate which possible characteristics could be considered when selecting exercises for learners and how humans adapt exercise selection to learner personality and performance, so that an Intelligent Tutoring System (ITS) can tailor exercise difficulty to these characteristics. In six focus groups, we had discussed which characteristics of the learner and the exercise could be important for exercise selection. With participants playing the role of the system, we then showed validated stories conveying learner personality traits of Conscientiousness, Self-Esteem and Emotional Stability at high and low levels and an indication of their previous performance on a simple mathematics exercise. Participants were shown an example of the kind of exercises which would be given to learners of varying difficulty and asked to select the exercise which they thought the learner should do next. We observed that participants responded based on the personalities of the learners presented as well as their past performances, with learners high in each trait being given a slightly more difficult exercise than learners low in that trait and learners who performed better being given more difficult exercises than learners who had performed poorly.

Keywords
Learning; Exercise Selection; Adaptation; Personality; Performance

1. INTRODUCTION

The personality of an individual plays a major role in determining how an individual responds to environmental situations [3, 13] and subsequently influences the decision making process of the individual. In the area of task selection in the learning domain, several characteristics such as past performance, cognitive load and support [2, 4, 5, 20] have determined which next task to give learners. However, the use of personality as a learner characteristic has been relatively unexplored; few works have used personality in selecting tasks for learners.

Learning tailored to individual characteristics has gained relevance in recent times [12, 17, 20, 11]. This learning process has progressed from the use of a fixed predefined pattern of learning tasks for all learners, resulting in better learning outcomes. There is also evidence that certain personality characteristics strengthen or reduce the effect of interest; for example, initiative and persistence are two aspects of action control that independently affect effort expenditure [19]. Disengagement interacts with interest: students who have the skill to uncouple a learning intention from an action plan are more affected by low interest than students who lack this skill [1]. [9] found that an individual’s learning orientation, and therefore their approach to learning, is partially determined by their personality. A deep approach to learning was positively associated with extraversion and openness to experience, while a surface approach was positively related to emotional stability and agreeableness. A strategic approach correlated positively with extraversion and conscientiousness and negatively with emotional stability. It is also established that there is a relationship between personality types and/or traits of the learners and their academic success in schools [18]. Therefore, personality should be taken into account when implementing Intelligent Tutoring Systems (ITS) for task selection and not just performance and cognitive load alone to produce better learning outcomes.

As a result of the evident relationships between personality and learning as shown in the above reviews, several adaptations to personality have evolved e.g. [11] adapted linguistic style to personality and [8, 7] adapted feedback to learner personality to improve motivation. Additionally, adaptive learning systems have adapted course and exercise sequencing in lessons to student progress [10].

As defined by [21], a focus group is a group interview which seeks to generate primarily qualitative data by capitalizing on the interaction that occurs within the group setting. Focus groups are usually centred on specific topics. [14] wrote that information saturation can be reached usually after discussions with about six groups. There is
2. FOCUS GROUP DESIGN

We conducted six focus groups (FG) because we had three personality traits to explore: Self-Esteem (confidence in one’s own worth or abilities), Emotional Stability (the inverse of neuroticism; being generally calm and less reactive to stress) and conscientiousness (how hard working, careful and thorough one is). We chose these traits to investigate as they seem the most applicable to the learning domain. We ran two focus groups for each trait; FG1A & FG1B investigated conscientiousness, FG2A & FG2B investigated Self-Esteem and FG3A & FG3B investigated emotional stability.

2.1 Participants

As the focus of this study is computer adaptation to learners, we decided to select participants from the computing and learning domain. Participants were recruited from students taking a course in the department of Computing Science at the University of Aberdeen. A total of 33 students participated in the focus groups, including postgraduates and undergraduate students. Table 1 shows the demographic information of participants. Participants were informed that their participation was voluntary and is not a requirement for the course. The focus groups were organised at the University of Aberdeen. A total of 33 students participated in the focus groups, including postgraduates and undergraduate students.

Table 1: Composition of Focus Groups

<table>
<thead>
<tr>
<th>Focus Group</th>
<th>Number of participants</th>
<th>Males</th>
<th>Females</th>
<th>Personality Trait</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>FG1A</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>Conscientiousness</td>
<td>Postgraduate students</td>
</tr>
<tr>
<td>FG1B</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>Conscientiousness</td>
<td>Postgraduate students</td>
</tr>
<tr>
<td>FG2A</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>Self-esteem</td>
<td>Undergraduate students</td>
</tr>
<tr>
<td>FG2B</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>Self-esteem</td>
<td>Undergraduate students</td>
</tr>
<tr>
<td>FG3A</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>Emotional stability</td>
<td>Undergraduate students</td>
</tr>
<tr>
<td>FG3B</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>Emotional stability</td>
<td>Undergraduate students</td>
</tr>
</tbody>
</table>

2.2 Research Questions

The focus groups were designed to answer three main research questions:

1. What do we need to know about a learner (learner characteristic) before giving them the next exercise to do?
2. What do we need to know about exercises (exercise characteristics) to know which one to pick next for the learner?
3. What next exercise should be selected for learners with different personalities and performances?

2.3 Procedure

The focus groups began with introduction of participants to each other. Participants were then told that the purpose for the focus groups was to discuss how an e-learning system could automatically adapt exercise selection to different types of learner characteristics. Information sheets and consent forms were distributed amongst participants and the opportunity was given to ask any questions.

We followed a semi structured approach. To answer research question 1, participants were asked: “Which learner characteristics of the learners do you think matter when deciding on a next exercise to give to a learner?” To answer research question 2, participants were asked “What do you need to know about the exercises to determine which one to pick next?”

To answer research question 3, participants were shown two stories depicting high and low self-esteem, as shown in Table 2.

Table 2: Stories depicting High and Low Self Esteem

<table>
<thead>
<tr>
<th>SE Level</th>
<th>Story</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Kate is a student who is confident about her abilities. She is satisfied about the way she looks and feels good about herself. She thinks she is as smart as others and believes that others admire and respect her. She feels that she has a good understanding of things.</td>
</tr>
<tr>
<td>Low</td>
<td>Nancy is a student who worries about the impression she makes and whether she is regarded as a success or a failure. She feels like she is not doing well and she believes she cannot understand the things she reads. Nancy thinks she is unattractive and is displeased with herself. She feels inferior to others.</td>
</tr>
</tbody>
</table>

The participants were then asked to place themselves in the role of the learner’s teacher, and pick the difficulty of next exercise for the learner to do. Participants could choose from: slightly more difficult, much more difficult, the same difficulty, slightly easier or much easier.

This procedure was repeated four times for each past performance level for the two learners.
2.4 Ethical Consideration

Consent forms and information sheets were distributed amongst the participants at the start of each focus group and participants were told that any material produced in the group may be used for publication but will be fully anonymised. An audio recording of the sessions was taken with the consent of the participants and notes were made. Taking part in the focus group was voluntary and participants were informed that they were allowed to withdraw from the focus group at any time and for any reason.

2.5 Materials

The materials used to conduct the focus groups were:

- Trait stories expressing personality traits (conscientiousness, self-esteem and emotional stability) at high and low levels
- Exercise card showing a sample of exercise for learners to do
- Performance card showing how the learner performed in previous exercises

3. RESULTS

The focus groups gave the opportunity to discuss some of the characteristics used previously for exercise selection and explore any other characteristics that should be considered. For the purpose of clarity, the results are arranged by responses to the research questions in the order outlined below. Participants answered the questions based on the materials presented to them. We present the results for the answers to questions 1 and 2 for Conscientiousness personality trait (FG1A and FG1B), Self-Esteem personality trait (FG2A and FG2B) and the emotional stability personality trait (FG3A and FG3B) and then question 3 for the same traits listed above. The results from the answers to question 3 spans through the four different conditions: *did well, just passed, just failed and did badly* at high and low levels. These conditions describe the past performances of the learners.

### 3.1 Q1: Learner Characteristics

Table 3 shows the results for Q1. FG1A were of the opinion that age and performance were appropriate to be considered as learners characteristics when selecting the next exercise for learners. FG1B choose past experience, learning style, competence and emotions (what the learner was feeling) as characteristics to be considered. FG2A mentioned learning styles, knowledge of the exercises, experience, interest in the learning process and past performance. FG2B mentioned the age of the learners and their personality. FG3A mentioned that past performance, learning style, experience, effort and personality should be considered as learner characteristics before giving exercises to learners. FG3B suggested the age of the learner, knowledge of the topic, culture of the learner in relation to the exercises, information available to the learner in relation to the exercise, level of confidence of the learner and experience with the subject area of the exercises ‘we could ask about their level of confidence in what we are giving them, like even if you have got a qualification in an area, you might not be very confident in it’.

Table 3 shows that participants think that age, past performance, past experience, learning style, knowledge and learner personality are most important overall.

### 3.2 Q2: Exercise Characteristics

This question was not asked in FG1A/FG1B due to time constraints. The learning content in relation with the age of the learners was considered by FG2A. In addition, how interesting the subjects are, difficulty levels and support was suggested. FG2B suggested feedback and the rules governing how the exercises should be done. Both FG3A and FG3B suggested difficulty levels of the exercises (*we should just scale out from a point onwards, basically go through the steps*).

FG3B also added that the relevance of the exercises to the learners (*they have to be relevant to the studies the learner is doing*). They also highlighted the form of presentation of the exercises, past experience with the exercises, support by way of examples available to the learners and the consistency in the structure of the exercises should be considered as exercise characteristics. FG3B also felt the the exercises should be of an appropriate difficulty for the learners (*You need to give them exercises that they could reasonably do, for example giving them an exercise of which they have no prior knowledge or experience could make them loose confidence*).

Table 3 shows that participants think that exercise difficulty and feedback given are the most relevant exercise characteristics overall.

### 3.3 Q3: Conscientiousness (FG1A and FG1B)

#### 3.3.1 Did well condition

FG1A choose a more difficult exercise for high conscientiousness learners in this condition. A slightly harder exercise was chosen by FG1B while for learners for low conscientiousness, a slightly harder exercise was chosen by FG1A. FG1B provided no results for the *did well* performance condition.

#### 3.3.2 Just Passed condition

For the high level of this condition, FG1B selected a mixture of more difficult and easier exercises while some of them
Table 3: Identified important learner and exercise characteristics from all focus groups. Focus group where characteristics were identified marked with X.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Conscientiousness</th>
<th>Self-Esteem</th>
<th>Emotional Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FG1A</td>
<td>FG1B</td>
<td>FG2A</td>
</tr>
<tr>
<td>Learner (Q1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>past performance</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>age</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>learning style</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>past experience</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>emotions</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>interest</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>knowledge</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>personality</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>confidence</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>culture</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>competence</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>effort</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Exercise (Q2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time to complete</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>relevance</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>presentation</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>difficulty</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>support</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>feedback</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ground rules</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

choose same level of difficulty (‘she is used to succeeding, give her same because if she has easier, she will know and feel bad’).

For the low level of this condition, FG1B selected same exercise with support. FG1A suggested that different exercises should be given the learner altogether but they should be with the same level of difficulty. They were of the opinion that this will show whether the learner passed by chance or not.

### 3.3.3 Just failed condition

For the high conscientiousness learner, some participants in FG1B suggested an easier exercise while some choose the same difficulty level exercise. For the low level conscientiousness learner, we had about half of the participants in FG1A and FG1B selecting easier exercises for the learner while the other half selected same level difficulty. A participant in FG1A suggested that a mixture of easier and same level difficulty exercises should be used.

### 3.3.4 Failed badly condition

For the high conscientious learner, FG1B selected same difficulty exercises while for the low conscientious learner, FG1A suggested that we make the exercises to be of the same level of difficulty but make sure the exercises are not exactly the same as the last one but different altogether. FG1B selected an easier exercise and also mentioned the need to change the approach and method of delivering the learning content.

### 3.4 Q3: Self-Esteem (FG2A and FG2B)

#### 3.4.1 Did well condition

For the high self-esteem learners, FG2A and FG2B selected a more difficult level exercise. Also, FG2A and FG2B selected a slightly more difficult level exercise for learners with low self-esteem.

#### 3.4.2 Just passed condition

Learners with high self-esteem were given the same difficulty level exercises by FG2A. FG2B selected a slightly difficult exercise (‘Kate seems like someone who likes a challenge, and maybe the exercises where not challenging enough, so increase the difficulty slightly’). For learners with low self-esteem, FG2A selected a slightly more difficult exercise, to ‘increase difficulty slightly so she does not get really bored’. In FG2B, 3 participants selected same difficulty levels and just 1 selected an easier exercise.

#### 3.4.3 Just failed condition

For high self-esteem, participants in FG2A decided on the same exercise while participants in FG2B selected an easier difficulty level. On the other hand for low self-esteem learners, both FG2A and FG2B selected exercises with same level of difficulty. They were of the opinion that selecting an easier exercise for this learner will further reduce the already low self-esteem of the learner.

#### 3.4.4 Failed badly condition

FG2A selected an easier exercise for the learner with high self-esteem while FG2B selected same difficulty level. For the low self-esteem learner, both FG2A and FG2B selected an easier exercise for the learner (‘if she got for example only one right, she has to go back and learn the principles all over again’).

### 3.5 Q3: Emotional Stability (FG3A and FG3B)

#### 3.5.1 Did well condition

All participants on both FG3A and FG3B selected a more difficult exercise for the high emotional stability learner.
Table 4: Difficulty of selected exercise for all personality traits and learner performance levels. Key: \( \bowtie \) More difficult, \( \land \) Slightly more difficult, \( \lor \) Same difficulty, \( \lor \) Easier, NR No Results

<table>
<thead>
<tr>
<th>Trait performance</th>
<th>Conscientiousness</th>
<th>Self-Esteem</th>
<th>Emotional Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FG1A</td>
<td>FG1B</td>
<td>FG1A</td>
</tr>
<tr>
<td>Did well</td>
<td>( \bowtie )</td>
<td>( \land )</td>
<td>( \land )</td>
</tr>
<tr>
<td>Just passed</td>
<td>NR</td>
<td>( \land ) &amp; = &amp; ( \lor )</td>
<td>( \land ) =</td>
</tr>
<tr>
<td>Just failed</td>
<td>NR</td>
<td>( \lor ) &amp; =</td>
<td>( \lor ) &amp; =</td>
</tr>
<tr>
<td>Failed badly</td>
<td>NR =</td>
<td>NR =</td>
<td>NR =</td>
</tr>
</tbody>
</table>

Also, FG3A and FG3B selected a slightly more difficult exercise for the low neurotic learner.

### 3.5.2 Just passed condition

For the high emotional stability learner, most of the participants in FG3A and all the participants in FG3B choose same level of difficulty (‘seems like a similar but next level of difficulty because the big important thing about her is that she remains calm under pressure and she probably did the best of her ability’), (‘should try to push her a little to the next level beyond her present ability and make her smarter, she won’t get irritated anyway’). Two participants in FG3A suggested a slightly harder exercise.

For the low emotional stability learner, participants in FG3A selected a slightly easier exercise, while participants in FG3B selected same exercises (‘same level as we could not give anything harder because she would be frustrated’), (‘easier because if Tina sees that she is able to pass, she will be motivated as such a person needs to be shown she is capable of passing a test’).

### 3.5.3 Just failed condition

FG3A and FG3B selected the same level of difficulty for learners with high emotional stability personality trait (‘knowing the kind of person that she is she stuff like having a bad day wouldn’t influence her that much, so similar level of difficulty should be given to her’).

For the low emotional stability learner, FG3A selected a slightly easier exercise (‘slightly easier because if she notices that it is much easier, it will make her feel bad about herself’) with some participants in this group suggesting a combination of easy and difficult exercises for the learner.

### 3.5.4 Failed badly condition

Both FG3A and FG3B selected easier exercises for learners with high emotional stability that performed very badly and for learner with low emotional stability that performed very badly, an easier exercise was also selected. Interestingly, FG3B agreed that the mood of ‘Tina’ (the learner with low emotional stability) should be taken into account more than ‘Emily’: (‘compared to Emily, you need to consider her mood, it could influence her learning’).

Learners high in all traits were given a more difficult exercise than learners who were low in the traits, who were only given a slightly more difficult exercise. For the ‘just passed’ condition, most of the participants choose the same difficulty level for the learners although for low self-esteem learner, a slightly more difficult exercise was selected so as to boost their self-esteem and motivation as agreed by the participants. Again, there was a trend towards learners high in the trait being given slightly more difficult exercises than learners low in the trait. For the ‘just failed’ condition, most of the participants selected the same exercises, however easier exercises were mostly selected for learners with low conscientiousness and learners with low emotional stability. An easier exercise was selected for almost all the learners that failed badly except for high conscientious and high self-esteem learners who were given exercises of the same difficulty level because the participants were of the opinion that giving them an easier exercise would be demotivating to them.

It seems therefore that, as expected, the performance of the learner is the primary adaptation characteristic for adaptation. However, we found important differences in these adaptations when the personality of the learner was considered. For each of Emotional Stability, Conscientiousness and Self-Esteem, participants thought that learners high in these traits should be given harder exercises than learners low in these traits. This suggests that researchers into intelligent tutoring systems should take learner personality into account when designing exercise selection algorithms. Future empirical studies will investigate (1) how exercise selection can be adapted to learner performance and personality (an initial study is reported in [16]), and (2) the effectiveness of such adaptations in keeping learners motivated and increasing learning outcomes. More information about the project as a whole can be found in [15].

### 5. ACKNOWLEDGMENTS

We thank the Niger Delta Development Commission (NDDC) for sponsoring the PhD of the first author.

### 6. REFERENCES


Exploring the Impact of Extroversion on the Selection of Learning Materials

Manal Alhathli  
University of Aberdeen  
Aberdeen, UK  
r01maea@abdn.ac.uk

Judith Masthoff  
University of Aberdeen  
Aberdeen, UK  
j.masthoff@abdn.ac.uk

Advaith Siddharthan  
University of Aberdeen  
Aberdeen, UK  
advaith@abdn.ac.uk

ABSTRACT

The Internet provides access to many learning materials that could complement class room teaching. An educational recommender system can aid learners to find learning materials most suitable to them. The best learning materials will depend on learner characteristics. This paper investigates the influence of learner personality. In particular, it describes a study in the language learning domain that explores the relation between learners’ extroversion and the extent to which learning materials are perceived to be enjoyable and to increase their confidence and skills. We found positive correlations between extroversion and these criteria for social and active learning materials.

Keywords
Personality, learning material, educational recommender, personalization

1. INTRODUCTION

E-learning has been gaining in importance, as also seen by the plethora of massive open online courses. In addition to established courses, there is a vast quantity of learning materials available on-line, such as YouTube videos. For learners, it can be quite difficult to find the materials best for them. Educational recommender systems may help to solve this problem. To be effective, such systems need to personalize their recommendations to learner characteristics, to ensure recommended materials suit individual users with different needs and requirements.

Our research aims to discover if a recommender system that incorporates learners’ psychological traits in it decision making would prove more effective than current applications. Therefore, this paper’s study aims to find evidence of whether a learner’s personality, in particular their degree of extroversion, should have an impact on the selection of learning materials. This work will feed into future work on constructing adaptive educational recommendation mechanisms specifically tailored to learners’ personalities.

2. RELATED WORK

2.1 Educational recommender systems

Educational recommender systems aim to improve the learning process by recommending appropriate courses, topics, peers, or learning materials (see [14] for an overview). We are particularly interested in recommendation of learning materials. For example, [13] have developed an on-line personalized English learning recommender system that provides reading materials for English Second Language learners. The educational recommender of [2] provides useful information and pages from the Internet for observed gaps in learners’ knowledge. The ISIS system recommends learning activities (as part of a Psychology course), providing navigation support in self-organized learning networks [8]. The educational recommender of [28] recommends learning documents from a set provided by the teacher and also discusses extending this to recommend text books. The educational recommender of [11] makes suggestions about similar materials based on learners’ ratings to enhance e-learning performance. The CoFind system [9] recommends learning materials based on folksonomies. The Altered Vista system [21] and QSIA system [20] recommend learning resources to members of a learning community.

2.2 Personality in learning

Several publications have documented five dimensions (i.e. extroversion, openness, conscientiousness, agreeableness and neuroticism) as underpinning the basic characteristics of personality, providing practical methods to appraise individuals [15]. Several studies have shown that there is a strong relationship between personality and academic performance [23]. For example, Openness to Experience was associated positively with academic performance [1]. Extroversion shows a positive correlation with participation in academic activities such as seminars [10] and low positive correlation with final exam grades [3]. Conscientiousness shows a strong positive correlation with final exam grades [3]. Furthermore, creativity in learning has been positively correlated with extroversion [22]. Another positive correlation can be found between four personality traits (extroversion, agreeableness, conscientiousness, openness) and a student’s motivation to attend college [4]. Given the clear evidence that personality influence learning, we would like to investigate how educational recommender systems can take personality into account.

There has been a strong tradition in e-learning to personalize interactive instruction systems to learners, leading to for example Intelligent Tutoring Systems. However, most
of these systems adapt to other learner characteristics, for example to performance, affective state, and learning styles (which differ from personality). Research on adapting e-learning systems to personality has been more limited. Dennis et al [7] investigated adapting learner performance feedback and emotional support to the Big5 personality traits. Okpo et al [19, 18] investigated adapting exercise difficulty to learner self-esteem. Robison et al [24] investigated the role of personality when giving different feedback types to learners. Del Soldato and Du Boulay [5] describe a motivational planner that can adapt its tactics to learner confidence. Nunes [17] describes a recommender that recommends compatible learners to work with based on Big5 personality traits. To the best of our knowledge, there is no previous work on recommending learning materials based on personality.

3. STUDY: IMPACT OF EXTROVERSION

3.1 Study Design

The study investigates which types of learning material are best for those with different personalities, in particular extroversion. There were three parts to the study. The first section gathered basic demographic information from participants and contained a short personality test. The second part asked participants to rate the learning materials for “John”, a fictional foreign language learner, who was described as having a similar personality to the participant. The final part asked participants to pick the best learning material.

3.1.1 Participants

The study was administered as a questionnaire on Mechanical Turk [16]. We included a Cloze Test [29] for English fluency to ensure that workers possessed enough literacy skills to understand the language based nature of the task. Participants had to have an acceptance rate of 90%, be based in the US and pass the fluency test in order to be eligible for the study. There were 50 participants (14 female, 35 male, 1 non-disclosed; 9 aged 18-25, 28 aged 26-40, 13 aged 41-65).

3.1.2 Materials

Foreign language learning was chosen as the domain, as many learning materials for it exist and it lends itself easily for different types of learning materials. Food ordering in a restaurant was chosen as the topic, as this is very popular in language courses. We used 7 learning materials, which a short textual description provided to participants (see Table 1). Learning materials were intended to be either passive or active, and individual or social. This was validated during the study by participants rating the extent to which the material involved John in active participation (active), and the extent to which it involved John in social interaction (social).

3.1.3 Variables

The independent variables are: the personality of ‘John’ (which matched the personality of the participant) focusing on extroversion, whether the learning material was active or passive, and whether the learning material was individual or social.

<table>
<thead>
<tr>
<th>ID</th>
<th>Learning Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>In this learning material, John will participate in an on-line spoken dialogue with a fellow learner about ordering food. John will play the role of the customer and the fellow learner the role of the waiter.</td>
</tr>
<tr>
<td>2</td>
<td>In this learning material, John will participate in an on-line spoken dialogue with a native speaker about ordering food. John will play the role of the customer and the native speaker the role of the waiter.</td>
</tr>
<tr>
<td>3</td>
<td>In this learning material, John will participate in an on-line spoken dialogue with a virtual agent (computer) about ordering food. John will play the role of the customer and the virtual agent (computer) the role of the waiter.</td>
</tr>
<tr>
<td>4</td>
<td>In this learning material, John will view a video about two native speakers having a dialogue in a restaurant. Next, the dialogue will be translated into John’s own language.</td>
</tr>
<tr>
<td>5</td>
<td>In this learning material, John will view a video about two other learners having a dialogue about ordering food in a restaurant.</td>
</tr>
<tr>
<td>6</td>
<td>In this learning material, John will view a video showing two other learners having a dialogue about ordering food in a restaurant. John can provide spoken feedback to the learners on their performance.</td>
</tr>
<tr>
<td>7</td>
<td>In this learning material, John will practice the food ordering vocabulary using multiple choice exercises.</td>
</tr>
</tbody>
</table>

The dependent variables are: the extent to which participants felt the learning material is enjoyable for John, increases John’s confidence in the language, improves John’s language skills, and the most preferred (‘best’) learning material to use for John. We will abbreviate the three ratings to enjoyable, confidence and skills below. Each of these ratings was given on a 5 point Likert-scale, from “not at all” to “a lot”.

3.1.4 Procedure

Participants first completed the English fluency test. If they passed, they provided their demographics and took a short personality test for the Five-Factor Model (FFM) [12], using Personality Sliders, a newly developed personality test [26]. For each trait from the FFM, participants were shown two stories (developed by [6]), one depicting a person that was low for that trait and the other depicting someone who was high. Participants used a slider to indicate which person they were most like, resulting in a value for each trait between 18 and 162. These are validated as accurately measuring the FFM [27].

On the next screen, participants were introduced to “John”, who has a similar personality to them. They were told that John is learning a foreign language and has just attended a class on ordering food in a restaurant. Next, they rated each learning material in turn in random order using the 5 scales (enjoyable, confidence, skills, active, social). Finally, they selected the learning material which was best for John.
3.2 Results

3.2.1 Types of learning materials

We first tested whether the learning materials did indeed match the individual versus social, and passive versus active distinctions we anticipated. Results are shown in Table 2. We conducted one sample t-tests to investigate whether the mean was significantly different from the mid-point of the scale (i.e. 3, see table for significance values). Based on this, four learning materials were found to be active (learning materials 1, 2, 3, 6), two passive (4 and 5), and one neither active nor passive (7). Three learning materials were found to be social (1, 2, 6), three individual (4, 5, 7), and one neither (3).

Table 2: Mean (stdev) of active and social ratings

<table>
<thead>
<tr>
<th>Learning material</th>
<th>Active</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.42 (.73)**</td>
<td>4.26 (.83)**</td>
</tr>
<tr>
<td>2</td>
<td>4.28 (.88)**</td>
<td>4.22 (.91)**</td>
</tr>
<tr>
<td>3</td>
<td>4.06 (.96)**</td>
<td>3.04 (1.21)</td>
</tr>
<tr>
<td>4</td>
<td>2.20 (1.09)**</td>
<td>1.96 (1.03)**</td>
</tr>
<tr>
<td>5</td>
<td>1.90 (1.00)**</td>
<td>1.80 (.99)**</td>
</tr>
<tr>
<td>6</td>
<td>3.52 (1.02)**</td>
<td>3.44 (.99)**</td>
</tr>
<tr>
<td>7</td>
<td>2.92 (1.23)</td>
<td>1.98 (1.17)**</td>
</tr>
</tbody>
</table>

**p<.01, ***p<.001

3.2.2 Extroversion and learning material ratings

First, we investigated the Pearson correlation between participants’ level of extroversion and their ratings for passive and active learning materials respectively. The results are shown in Tables 3 and 4. For passive learning materials, we found no significant correlations. For active learning materials, we found significant and positive correlations of extroversion with (1) the enjoyability of learning materials, (2) whether the learning material increased learner’s confidence, and (3) whether it increased learner’s language skills. These results may be explained by the fact that most active learning materials were social, whilst the passive ones were all individual.

Table 3: Correlations for passive learning materials

<table>
<thead>
<tr>
<th>Extroversion</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Language-skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.165</td>
<td>.068</td>
<td>.053</td>
</tr>
</tbody>
</table>

Table 4: Correlations for active learning materials

<table>
<thead>
<tr>
<th>Extroversion</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Language-skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.275**</td>
<td>.179*</td>
<td>.160*</td>
</tr>
</tbody>
</table>

*p<.05, **p<.01

Secondly, we investigated the correlations between participants’ level of extroversion and their ratings for individual and social learning materials respectively. The results are shown in Tables 5 and 6. For individual learning materials, we found no significant correlations. For social learning materials, we found significant and positive correlations of extroversion with (1) the enjoyability of learning materials, (2) whether the learning material increased learner’s confidence, and (3) whether it increased learner’s language skills.

These results can be explained by extroverts preferring social learning materials.

Table 5: Correlations for individual materials

<table>
<thead>
<tr>
<th>Extroversion</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Language-skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.158**</td>
<td>.011*</td>
<td>.017*</td>
</tr>
</tbody>
</table>

Table 6: Correlations for social learning materials

<table>
<thead>
<tr>
<th>Extroversion</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Language-skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.315**</td>
<td>.206*</td>
<td>.195*</td>
</tr>
</tbody>
</table>

Next, we split the participants into two groups, based on their score for extroversion: an extrovert group with an extroversion score above the mid point of the scale, and an introvert group with a score below the mid point of the scale. There were 17 extroverts and 33 introverts. Ratings per group for the different types of learning materials are provided in Tables 7-10.

Table 7: Mean (stdev) for social materials

<table>
<thead>
<tr>
<th>Extroverts</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.49 (.67)</td>
<td>3.86 (.83)</td>
<td>3.98 (.93)</td>
<td></td>
</tr>
<tr>
<td>Introverts</td>
<td>2.87 (1.12)</td>
<td>3.49 (1.02)</td>
<td>3.69 (.97)</td>
</tr>
</tbody>
</table>

Table 8: Mean (stdev) for individual materials

<table>
<thead>
<tr>
<th>Extroverts</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.04 (1.04)</td>
<td>3.04 (1.08)</td>
<td>3.16 (1.12)</td>
<td></td>
</tr>
<tr>
<td>Introverts</td>
<td>2.81 (1.24)</td>
<td>3.13 (1.20)</td>
<td>3.24 (1.12)</td>
</tr>
</tbody>
</table>

Table 9: Mean (stdev) for passive materials

<table>
<thead>
<tr>
<th>Extroverts</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.06 (.98)</td>
<td>3.06 (1.13)</td>
<td>3.03 (1.17)</td>
<td></td>
</tr>
<tr>
<td>Introverts</td>
<td>2.80 (1.27)</td>
<td>2.98 (1.25)</td>
<td>3.06 (1.15)</td>
</tr>
</tbody>
</table>

Table 10: Mean (stdev) for active materials

<table>
<thead>
<tr>
<th>Extroverts</th>
<th>Enjoyable</th>
<th>Confidence</th>
<th>Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.55 (.74)</td>
<td>3.84 (.86)</td>
<td>3.97 (.91)</td>
<td></td>
</tr>
<tr>
<td>Introverts</td>
<td>3.00 (1.11)</td>
<td>3.58 (1.02)</td>
<td>3.77 (.94)</td>
</tr>
</tbody>
</table>

Considering enjoyment, the extrovert group rated both active, passive, social and individual learning materials as more enjoyable than the introvert group, however the differences between extroverts and introverts are only significant for social and active materials (p<.05), in line with the correlation results.

Considering increasing the learner’s confidence, the extrovert group rated the social learning materials a lot higher than the individual ones. The introvert group also rated social materials more highly, but the difference was a lot smaller than for the extroverts. Interestingly, extroverts rated social materials higher than introverts, but rated individual materials lower, but this was not statistically significant. Looking at the difference between active and passive
learning materials, both extroverts and introverts rated the active materials a lot higher than the passive ones, and in both cases the extroverts had slightly higher ratings than the introverts (though not significant).

Considering increasing the learner’s language skills, the extrovert group rated the social learning materials a lot higher than the individual learning materials. The introvert group also rated social materials more highly, but the difference was a lot smaller than for the extroverts. Interestingly, extroverts rated social materials higher than introverts, but rated individual materials lower. Looking at the difference between active and passive learning materials, both extroverts and introverts rated the active materials a lot higher than the passive ones. The difference between extroverts and introverts for passive materials is negligible, whilst for active materials extroverts rated slightly (but not significantly) higher.

### 3.2.3 Extroversion and learning material selection

Figure 1 shows the selection of the best learning material among the two groups of introverts and extroverts. It can be clearly seen that the majority of the two groups find that social and active materials are the best to recommend to John (though interestingly, learning material 6 which is both social and active is completely absent). The second most selected learning materials in introverts were passive and individual. Interestingly learning materials 5 and 7 (both individual and neither active) were not selected at all in the extroverts' group, whilst they were selected in the introverts group. There also seems to be an interesting difference for learning material 3: this was the material where learners had a dialogue with a virtual agent (instead of with a human as in learning materials 1 and 2). This seems more popular with introverts than extroverts. More statistical analysis can be done here.

### 4. CONCLUSIONS AND FUTURE WORK

In this paper we presented a first study on how a computer can recommend learning materials to users and investigated in particular the effect of learners’ extroversion on their appreciation of learning materials on three criteria. We presented an initial analysis of the data, which showed that extroversion was weakly but positively correlated for active and social learning materials with learners’ rating of how much they thought a learner with a personality similar to their own would (1) enjoy the learning material, (2) increase their confidence through the learning material, (3) increase their skills through the learning material. Further statistical analysis showed significant differences between introverts and extroverts for enjoyment. Future work will extend the analysis of the data and will use the results of this study and follow-up studies to inform the design of adaptive recommendation algorithms.

This paper presented an indirect study, we did not measure actual enjoyment, actual increase in confidence and actual increase in skills, but perceptions of those. In a sense, therefore the study looked at learner preferences. Clearly there is more to an effective educational recommender system than learner preferences (as noted by [25]). We plan to run follow-on studies in a real learning environment to obtain more direct measures, also on learning gain. In addition, we plan to interview teachers to obtain their input.

This paper only investigated extroversion. We plan to conduct other studies to investigate other personality traits as well as their interaction with other learner characteristics (e.g. learners’ goals and interests, knowledge and performance, learning styles, age). In this study, we recruited participants via Amazon Mechanical Turk, meaning they all came from the US. It is possible that the preference for social and active learning materials is related to participants’ cultural backgrounds, and this can be investigated in future studies.

The study used the domain of learning a foreign language. This may have had an impact, as skills’ and confidence development in language learning may benefit more from social interactions. We plan to repeat the study in another learning domain. In this study, we only distinguished between active/passive and social/individual learning materials, as we assumed that those were the most important for the extroversion personality trait. The learning topic was the same for all learning materials, and difficulty of the learning materials was not considered. Other learning material characteristics (and their interactions) still need to be investigated. For example, it is possible that for openness to experience the novelty or diversity of learning material may be relevant (c.f. [30] for adaptations of recommender systems’ diversity to openness of experience).

Finally, we did not investigate different ways to present recommendations to learners. It is possible to use for example Top-N recommendations, presenting multiple ranked materials for learners to choose from, or to indicated the expected suitability of a learning material for instance by stars. An overview of this issue for recommenders in general can be found in [31], and more work on presentations and explanations of educational recommendations is required.
5. ACKNOWLEDGMENTS

We thank the Ministry of Higher Education in Saudi Arabia for sponsoring the PhD of the first author.

6. REFERENCES

[16] MT. Amazon mechanical turk.


PALE 2016 (Edited by M. Kravcik, O.C. Santos and J.G. Boticario, M. Bielikova) 41
Combining Content Analytics and Activity Tracking to Identify User Interests and Enable Knowledge Discovery

Andrii Vozniuk, María Jesús Rodríguez-Triana, Adrian Holzer, and Denis Gillet

REACT Group, EPFL, Station 9, 1015 Lausanne, Switzerland.
Email: {andrii.vozniuk, maria.rodrigueztriana, adrian.holzer, denis.gillet}@epfl.ch

ABSTRACT
Finding relevant content is one of the core activities of users interacting with a content repository, be it knowledge workers using an organizational knowledge management system at a workplace or self-regulated learners collaborating in a learning environment. Due to the number of content items stored in such repositories potentially reaching millions or more, and quickly increasing, for the user it can be challenging to find relevant content by browsing or relying on the available search engine.

In this paper, we propose to address the problem by providing content and people recommendations based on user interests, enabling relevant knowledge discovery. To build a user interests profile automatically, we propose an approach combining content analytics and activity tracking. We have implemented the recommender system in Graasp, a knowledge management system employed in educational and humanitarian domains. The conducted preliminary evaluation demonstrated an ability of the approach to identify interests relevant to the user and to recommend relevant content.

Categories and Subject Descriptors
K.3.1 [Computers and Education]: Computer Uses in Education; H.5.2 [Information interfaces and presentation]: User interfaces

Keywords
Learning Analytics, Educational Data Mining, Interests Mining, Knowledge Discovery, Recommender System, Content Analytics, Text Mining, Activity Tracking, Information Retrieval

1. INTRODUCTION
Knowledge plays an essential role in value creation in the post-industrial economy. Knowledge is acquired and enriched in learning, which often takes place at a workplace or in an educational setting. While learning, people interact with content as a knowledge medium, located in various content repositories. In an educational setting, such content repositories are usually learning environments, where both students and teachers interact with the content found there. Teachers would regularly interact with the content, when preparing a course while students - when following a course or just collaborating with peers.

When working on a course in a learning environment, teachers enrich the system with relevant materials including text files, web links, videos, audio recordings coming from their device or the cloud. Other teachers can benefit from content already available in the platform when preparing their courses. Moreover, it may also be beneficial for the students to have access to the content that is relevant to their interests, but which the teacher did not directly include into her course [25]. In the case of learning environments with a vast number of content items, it may be hard for the user to find content items corresponding to her interests.

To address the mentioned issues, we propose to employ a recommender system that combines the content analytics, activity tracking, and information retrieval techniques to (1) build the user interests profile and afterward (2) to suggest content relevant to the user and users with similar interests enabling knowledge discovery. To perform the recommendation, first, for each item available in the content repository, we employ natural language processing techniques to identify a set of concepts related to the content in a similar way how humans would do it. Relying on high-level concepts instead of specific words present in the text when constructing user interests profile and afterward finding similar items, allows to identify the content that covers the same high-level concepts even if the specific words used in it are different.

Next, we analyse the interactions of the users with the content items based on available user activity recordings and aggregate the concepts in the content that the user interacted with building in this way the user interest profile. Finally, we use information retrieval techniques to recommend to the user relevant content based on the similarity between the concepts in the content and concepts identified as user interests. In the same way, our approach allows finding relevant users based on the determined interests similarity. Our approach puts the user in control of her interests profile and allows to adjust the interests by removing concepts if necessary, as in the case when the user is not interested at the moment in some of the identified concepts.

To evaluate the usefulness and the performance of the algorithm, we have implemented the approach in Graasp, a knowledge management system used in educational [2] and humanitarian [24] settings. Afterward, we have evaluated the approach with teachers, identifying their interests and providing them with recommendations.

This paper describes the algorithm used, the implementation details of the approach in Graasp and the evaluation of the approach with users. The structure of this paper is as follows. First, Section 2 reviews some of the relevant approaches to content analytics, activity tracking, and knowledge discovery. Afterward, Section 3 explains our approach to constructing user interests profile and demonstrates how
we make recommendations based on the interests. Section 4 illustrates an implementation of the proposal, while Section 5 talks about the evaluation methodology and the results. Finally, Section 6 presents the conclusions and highlights directions for the future work.

2. RELATED WORK

In this section, we review relevant work from the domains of content analytics, activity tracking, user interests mining and take a look at notable systems supporting knowledge discovery.

2.1 Content Analytics and Activity Tracking

Content Analytics. Content analytics allows the machine to gain an understanding of the content, similarly to how a human would do it by, among others, extracting the main topics, concepts, and entities present in the content. Kovacic et al. did an extensive overview of content analytics as one of the often employed techniques in the domain of learning analytics [12]. For instance, in the line of our work, Bosnic et al. proposed to use automatic extraction of keywords from textual content as a foundation for content recommendations [3]. It is worth noting that existing papers focus mainly on analysis of textual content [12], while recent progress in the understanding of multimedia formats, such as object recognition in images or videos [13], or speech recognition allow broadening the scope of content analytics from purely textual information to the various multimedia formats.

Understanding the content alone is not sufficient for understanding the learning since, according to Moore, learner-content interaction is a defining characteristic of education [14]. Moore argues that such learner-content interaction is necessary to happen for the education to take place since "it is the process of intellectually interacting with content that results in changes in the learner’s understanding, the learner’s perspective, or the cognitive structures of the learner’s mind" [14]. Recognising the importance of the interaction, below we consider approaches to capturing and persisting the interactions through activity tracking.

Activity Tracking. User activities tracked by a learning platform is a common data source in the field of learning analytics [18, 19]. Usually, a learning management system or a learning environment have a logging infrastructure in place that records how the user interacts with the platform [18]. The more modern educational platforms support a structured representation of user activities using well-defined formats including ActivityStreams used in [23], xAPI employed in [11] or IMS Caliper outlined in [19]. On a high-level, all these three formats record user-platform interactions in the form of the actor-verb-object triplet capturing who did what with what on the platform. However, on a more detailed level, each format captures additional aspects of the interaction. In the triplet, the verb indicates the type of interaction, for instance, the verb "accessed" would mean that the user viewed content, "downloaded" - downloaded the content and so on. Having a common set of verbs with a well-defined meaning is critical for being able to benefit from user interactions captured by several platforms [11].

Combining Both. While there is a considerable number of studies employing content analytics or relying on interaction analysis, the number of studies combining both is still somehow limited even taking into account that it is considered a promising direction [18, 12]. One noticeable recent proposal combining the both approaches is by Kim et. al. [10] where they use content analytics and recorded interaction data to understand better how students learn with video and eventually to improve their experience, for instance by explaining better the identified confusing topics. Following these recommendations, we consider the combination of both content analytics and activity tracking as a core part of our proposal.

2.2 Mining User Interests

The obtained user interests can be used for different purposes, including privacy awareness and recommendations. Harrou et. al. proposed in [9] to employ a content analysis of the files located on Google Drive of a user to understand the topics, concepts, and entities relevant to the user. They used the obtained information with the goal to improve the user awareness through a new permissions model called Far-reaching Insights. This model informs the user about the insights that third-party applications can derive about her based on the accessible Drive data given the requested permissions are granted. In our approach, we want to explore how identified interests can be used to provide the user with relevant content. In the following subsection, we review some of the systems enabling knowledge discovery with such recommendations.

2.3 Knowledge Discovery Systems

Klamma et al. have formulated a set of requirements for a collaborative adaptive learning platform [16]. One of the requirements is "Support for personalized learning resource delivery through an intelligent adaptive engine, being able to connect people to the right knowledge and deliver quality learning resources that are tailored to the learner’s preferences and learning goals." [16]. Learning platforms often integrate such engine in a form of a recommender system. Drachsler et. al. have conducted an extensive review of 82 recommender systems used to support learning in [5]. Below, we take a look at several proposals, particularly relevant to our approach.

Zaldivar et. al. address in [25] the problem of discovering by the instructors relevant learning resources used by students when learning, that are not part of the materials provided by the instructor but still can be beneficial for the students. In their approach, the authors record the web pages that students visit and perform a lexical analysis of the page content. Afterward, they apply information retrieval techniques to identify the online content (webpages) that are the most similar to the content provided by the instructors as part of the course.

In [6] El Helou et. al. proposed a recommender system that considers user interactions with content items to construct a user-content associations graph. After the graph is built, the system applies a ranking algorithm to provide the user with personalized recommendations of relevant actors, activity spaces and knowledge assets taking into account the context.

Motivated by the presented approaches, in the next section, we propose to employ a recommender system that combines content analysis, activity tracking to identify user interests and information retrieval techniques to suggest relevant content and people.
3. INTERESTS-BASED RECOMMENDER

In this section, we explain how our approach works by first automatically identifying user interests and after using the interests to obtain relevant content and people.

3.1 Identifying User Interests

To identify user interests, our system needs, first, to understand the concepts covered in the content. Second, it requires recorded user activities to know how the user interacts with the content items. Having both the concepts and the activities, the system can construct the user interests profile. Below, we explain each component of the approach.

Content Analytics. Content can be available in multiple formats, and a data processing pipeline needs to be built to extract concepts from the content and after store them in an index for further use. A general representation of the key steps of the pipeline is shown on Figure 1 where different types of content may go through different processing steps to obtain the concepts.

On the first step, textual content is extracted from the stored items. In the second step, the content analysis is performed. For the content analysis, we considered using named entity recognition (NER), concept extraction, and topic modelling. Since NER picks entities only from the words present in the text, using such entities for recommendations would limit the discovery only to the content containing them directly. Differently, high-level concepts allow identifying relevant content even if the specific words used in it are different. When we applied topic modelling to real data, the identified topics having a high level of abstraction did not seem to capture well the content particularities. For these reasons, we use a set of concepts to describe the content. Finally, on the third step the extracted content and the concepts are tokenized and put into a searchable index so that they can be used on the recommendation step.

Activity Tracking. Our approach requires recording user-content interactions, namely the triplet user-verb-object. We consider different types of interactions as a manifestation of different interest strength. For instance, intuitively when a user downloads the content it manifests a stronger interest in the content; the identifier of the resource the user has to capture the user identifier, the verb indicating the type of interaction and, the identifier of the resource the user has interacted with.

Computing User Interests. As the user interacts with the content, the system aggregates the concepts identified in the content, weighting them according to the type of interaction and, the identifier of the resource the user has interacted with.

To identify user interests, our system needs, first, to understand the concepts covered in the content. Second, it requires recorded user activities to know how the user interacts with the content items. Having both the concepts and the activities, the system can construct the user interests profile. Below, we explain each component of the approach.

The aggregated concepts from the content, weighting them according to the type of interaction and, the identifier of the resource the user has interacted with. Hence, $U_{CI}$ is the relevance of the concepts $c_j$ for the user $u_i$; $w_v$ is the weight assigned to specific interaction type $v$ indicating how strongly specific action of the user expresses her interest in the content; $U_{A_{nm}}$ is the matrix capturing user-content interactions of type $v$; $DC_{m+1}$ contains relevance values for the item concepts.

While the formula presented above is suitable for computing the profile first time when the recommender is deployed, the profile does not need to be recomputed from scratch and can be updated incrementally. On every user-content interaction, we update in real-time the user concepts of interest based on the ones that were found in the content as follows:

$$U_{Conf}^{ua} = U_{Conf}^{ua} + w_v \cdot U_{A_{nm}} \cdot DC_{m+1}.$$  \hspace{1cm} (2)

where $U_{Conf}^{ua}$ is the vector of user concepts before the interaction and $U_{Conf}^{ua}$ - after the interaction; $U_{A_{nm}}$ is a matrix having 1 in position $(i, j)$ if the user $u_i$ has interaction of type $v$ with the content item $d_j$. All other elements are 0; and $DC_{m+1}$ contains relevance values for the item concepts.

Once the profile constructed, in the next section we explain how it can be used for recommendations.

3.2 Recommending Relevant Content and Users

Connecting right people with right knowledge is a possible way to improve knowledge sharing. We aim to improve knowledge discovery by facilitating connection creation between knowledge sources and users in need of knowledge. Knowledge sources can be individual content items or other users with similar interests possessing the knowledge. We propose an approach that can suggest 1) content relevant to users and 2) users with similar interest. Below, we present two main steps of our approach.

Step 1. Computing term weights with TF-IDF.

On the first step, we compute the relevance of specific terms (including concepts) for the content items by using a known information retrieval technique, namely term frequency - inverse document frequency (TF-IDF) as explained in [17]. When computing the weight, TF-IDF considers the frequency of the term inside of a document and its frequency in the whole corpus. In this way, for each content item we obtain a vector that contains weights of individual words or concepts $cw_{ci}$ present in the content:

$$cw_{ci} = tf_{ci} \cdot idf_{ci},$$  \hspace{1cm} (3)

where $tf_{ci}$ is the term frequency representing how often the term $ci$ appears in the document and $idf_{ci}$ is the inverse document frequency indicating how common is the term $ci$ in all documents.

Step 2. Scoring relevant items with cosine similarity.

To obtain for the user $u$ suggested content items or relevant users, we compute the relevance score for the item $d$ using a cosine similarity between the two vectors representing the user and the content:

$$S(u, d) = \frac{V(u) \cdot V(d)}{|V(u)||V(d)|},$$  \hspace{1cm} (4)

where $V(u)$ and $V(d)$ are the vectors containing weights
of the user terms and the document terms computed at Step 1; $V(u) \cdot V(d)$ is a scalar product of the two vectors; $|V(u)|$ and $|V(d)|$ are Euclidean norms of the vectors.

4. IMPLEMENTATION

To validate the feasibility of the approach and further evaluate it, we have implemented it in Graasp, a social media platform employed for knowledge management. Graasp supports uploading and storage of content from user devices or the cloud. Graasp provided extraction of text content from multiple file formats, and the activity logging infrastructure was already in place. Still, we needed to extend the platform to enable content analytics with concepts extraction, construction of the interests profile, and items recommendations with Elasticsearch\footnote{Elasticsearch Open Source Engine \url{https://github.com/elastic/elasticsearch}}. Below, we explain the architecture of the implemented solution.

4.1 Concept Extraction and Activity Tracking

Concept Extraction. The concepts extraction is done as soon as content is uploaded to Graasp. To extract concepts, we have implemented a processing pipeline presented on Figure 1. On the first step, the type of the content is identified, and Graasp tries to extract textual information when possible. For plain text files, it just reads the text content of the file. For binary text files including pdfs and

Figure 1: A possible pipeline architecture to extract concepts from diverse content types. Dotted lines mark the parts yet to be implemented in Graasp.

Figure 2: A schematic representation of the proposed approach. The system aggregates the concepts from the content as the user interacts with the content.
Microsoft Office formats, we use the textract library\(^2\). For images, Graasp tries to perform Optical Character Recognition and read the text presented on the image using the tesseract\(^3\) library. In the future, we foresee extracting text from Audio and Video files relying on Speech-To-Text technologies (shown with dotted lines on Figure 1) and obtaining concepts for images and videos with the help of visual recognition tools\([13]\), for instance using clarifai\(^4\). Once the text is available, we analyse its content, identifying the concepts present there. For this purpose, we concatenate the item name, the item description and the extracted content and, at the moment of writing, employ AlchemyAPI\(^5\) Concept Tagging to get the concepts. It is worth noting that our approach does not assume a specific concept identification technology, and AlchemyAPI was picked for the reasons of minimal administration and scalability. After the system identifies the concepts, it indexes them in Elasticsearch to-gether with the text content extracted before.

**Activity Tracking.** Graasp uses ActivityStreams format for capturing user activities on the platform. Some of the actions that the platforms records include access, download, rating, commenting, inviting members and, searching.

### 4.2 Interests and Recommendations

**Constructing Interests Profile.** Graasp continuously updates interests profile of the users as they interact with the content. Users interests are displayed next to their profile information as demonstrated on Figure 3. The user can adjust her profile by removing individual concepts by pressing the X button and in this way influence in real-time the content and users suggested by the recommender.

**Computing Recommendations.** In Graasp, we rely on Elasticsearch for computing recommendations whenever the user wants to see them. Elasticsearch is built on the Lucene\(^6\) text search engine that internally employs vector space model, TF-IDF, and cosine similarity when finding relevant items\(^7\), similarly as in our proposed approach described in Section 3.2. We assemble into a single search query all of the concepts from the user interests profile and, whenever present, the terms from the user description as on Figure 3 (1). We run this query against the name, description, content and, concepts fields of the items, assigning different boost weights for matches happening in different fields. The obtained results are presented to the user next to her profile as illustrated on Figure 3.

## 5. EVALUATION

To understand opinions regarding the approach and its performance when put into practice, we have conducted a preliminary evaluation of the approach implementation in Graasp with pre-service teachers. This section explains in more details the methodology used and the main outcomes.

### 5.1 Methodology

We have conducted a survey-based preliminary evaluation of the developed approach. Surveys are one of the common ways of evaluating recommender systems allowing to collect opinions regarding the system from multiple users in a reasonable timeframe\([7,21]\). Our goal was to validate if the approach, in general, is useful, if its implementation in Graasp is usable, as well as if the system can identify relevant interests and recommend relevant items.

**Participants.** We have conducted the survey with six participants of a workshop on inquiry-based learning for pre-service teachers in secondary education. During the workshop the participants registered in Graasp and carried out on the platform a set of activities during 2 hours. At the end of the session, we asked them to fill in the survey.

**Survey Structure.** Our survey had three parts\(^8\). The first part asked about general disposition towards the interests identification and the interests-based recommender. The second part was the System Usability Scale (SUS)\([4]\) evaluating the usability of the implemented system. We have selected SUS because of its understood interpretation and robustness\([1]\). In the third part, we evaluated the quality of the identified interests and recommendations. Two types of questions formed the survey. The first type was questions to indicate the level of agreement with specific statements, where we followed the 5-point Likert scale ranging from 1 - Strongly Disagree to 5 - Strongly Agree to obtain quantitative results. The second type was open questions where we asked the responders to provide us with qualitative feedback regarding the approach and its implementation.

### 5.2 Results

In this section we focus on the main outcomes of the evaluation. Complete survey results are available online\(^9\).

**Approach.** The users valued positively the idea of using their interests to guide the recommendations and to find other users with similar interests (mean Likert score \(\mu = 3.17\) and \(\mu = 3.33\) respectively). Besides, the fact of being aware of the inferred interests and the possibility of editing interests were well appreciated (\(\mu = 3.17\) and \(\mu = 3.33\) respectively). Although the quantitative analysis does not illustrate a high adoption by the users, during the workshop, the participants were keen on understanding how the interests were extracted and highlighted the novelty of the approach. Further details with the survey results may be found following the URL mentioned above.

**Usability.** In general, the participants were eager to use the recommender with a certain frequency (\(\mu = 3.33\)) and did not report major issues regarding complexity, inconsistency or difficulty of usage. Just one person considered that she would need technical support or previous background to use the recommender. According to the discussion with this person after the workshop, these answers were partially conditioned by the cognitive load due to the short time available to get used to the platform itself and to integrate all the ideas presented in the workshop. The quantitative results of the SUS questionnaire are also available on-line.

**Accuracy.** Despite the limited amount of traces collected due to the short time of the user interaction, the results point out that both the interests extracted and the recommendations, in general, were relevant (\(\mu = 3.17\)) and diverse (\(\mu = 3.50\)). It is noteworthy that when we asked the users to check how many relevant interests and recommendations

\(^2\)textract https://github.com/dbashford/textract
\(^3\)tesseract https://github.com/tesseract-ocr
\(^4\)Clarifai Library https://github.com/clarif.ai/
\(^5\)AlchemyAPI http://www.alchemyapi.com
\(^6\)Apache Lucene https://lucene.apache.org
\(^8\)Recommender Evaluation Survey https://goo.gl/Wes6uP
\(^9\)Evaluation results https://goo.gl/Wes6uP
Figure 3: (1) User interests in Graasp as identified by our approach. Suggested content (2) and suggested people (3) based on the user profile information and identified interests.

appeared in the top 10, we discovered two groups. While most of the users reported more that six relevant items, two users got less than two relevant items. We have looked into this case and identified the reason covered below.

Sensitivity to Inaccurate Concepts. In the case when an item that the user interacted with many times has concepts identified not accurately, these concepts appear on the top of user interests. We plan to mitigate this problem in the future by introducing a heuristic for not considering concepts with low relevance and by limiting the influence of a single item on the overall concept relevancy for the user. Our goal is to make sure that the identified concepts come from many items rather than from many visits to a single item with potentially misidentified concepts. Our expectation is that it will allow reducing the impact of faults in concepts identification on the user resulting user profile.

Privacy Implications. Right now, only the user can see her interests, but we consider putting in place a mechanism that will allow to make validated interests visible to other users of the platform and to make it possible to find the user based on her interests as it was proposed in [15]. However, based on the evaluation, while some of the participants were eager to make their interests visible, others were reluctant. Thus, it will be necessary to allow users configure the visibility of their interests to preserve their privacy, following the recommendations provided in codes of practice for learning analytics [20].

6. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a new approach to building user interests profile based on 1) content analytics providing the system with the concepts present in the content and 2) activity tracking allowing the system to know how the user interacted with the content. We have used the extracted interest concepts to recommend relevant content and people. Further on, we have implemented the proposed approach in Graasp, a knowledge management system. Graasp was used in a workshop to support teachers when building inquiry learning spaces for their students. Thus, we have evaluated the approach with the teachers, and the evaluation has demonstrated that the proposed approach can identify relevant user interests and recommend relevant content based on the identified interests. At the same time, the evaluation has unveiled sensitivity of the approach to inaccurately identified concepts that we plan to overcome in the future. While we draw our experience and motivation from the educational context, our contributions have a broad impact and can be applied for content repositories, where it is possible to obtain content analytics and track activities performed by the users (e.g., Google Drive and Dropbox).

Looking Outside. In this study, we analyzed the content and recorded the activities limited to the scope of the content repository. However, in the current technological landscape, the interactions are getting more distributed often spanning across multiple platforms. Studies suggest that combining data obtained from several platforms could allow creating a more accurate user interests profile [8]. In the future, we plan to extend the architecture of Graasp to aggregate the content and the interactions outside of the system.

Incorporating Relevance Scores. At the moment, when computing the similarity score for the recommended items we consider the fact of the concept presence but do not take into account the available concept relevance scores. Incorporating the relevance scores available for user interest concepts and content concepts when computing the user-content relevance score may lead to more relevant recommendations since it will promote the results with similar highly relevant concepts.

Recommender Adaptability. One potential downside of our approach could be related to its limited ability to react timely to change in the user interests reflected in her interactions. This happens since the concepts the user accumulated at some point through her interaction history maintain the same score indefinitely. One of the possible solutions to this problem is to introduce the forgetting function as suggested in [22], so that as time goes the concepts that are not encountered anymore get their relevance score reduced.

Substantial Evaluation. This paper presented a preliminary evaluation of the recommender to provide early feedback. We are planning to conduct an evaluation with more users that used Graasp for longer periods of time so that more activity traces are available. We also expect these users to have established expectations regarding their interests when interacting with the platform.
7. ACKNOWLEDGMENTS

This research is partially funded by the European Union in the context of the Go-Lab project (Grant Agreement no. 317601) under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D (FP7). This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

8. REFERENCES

ABSTRACT

Big Data and data technologies increasingly find their way into school education. Learning Analytics and Educational Data Mining are focal research areas. However, technical solutions often fail to meet the practical requirements of teachers or to really mirror human learning processes. The LEA’s BOX project aims at developing a practical web platform that hosts tools for a theory-based approach to Learning Analytics and that offers tools to open and negotiate learner models.

Keywords
Learning analytics, data visualization, open learner models.

1. INTRODUCTION

Using Learning analytics and educational data mining are more than recent buzz words in educational research: they signify one of the most promising developments in improving teaching and learning. While many attempts to enhance learning with mere technology failed in the past, making sense of a large amount of data collected over a long period of time and conveying it to teachers in a suitable form is indeed the area where computers and technology can add value for future classrooms. However, reasoning about data, and in particular learning-related data, is not trivial and requires a robust foundation of well-elaborated psycho-pedagogical theories.

The fundamental idea of learning analytics is not new. In essence, the aim is using as much information about learners as possible to understand the meaning of the data in terms of the learners’ strengths, abilities, knowledge, weakness, learning progress, attitudes, and social networks with the final goal of providing the best and most appropriate personalized support. Thus, the concept of learning analytics is quite similar to the idea of formative assessment. “Good” teachers of all time have strived to achieve exactly this goal. However, collecting, storing, interpreting, and aggregating information about learners that originates from a school year, or even in a lifelong learning sense) requires smart technology. To analyse this vast amount of data, give it educational meaning, visualize the results, represent the learner in a holistic and fair manner, and provide appropriate feedback, teachers need to be equipped with the appropriate technology. With that regard, a substantial body of research work and tools already exist. This project aims to continue and enrich on-going developments and facilitate the broad use of learning analytics in the “real educational world.

2. LEA’s BOX

LEA’s BOX (www.leas-box.eu) is a project, funded under the EU’s Seventh Framework Programme and stands for a practical LEarning Analytics tool Box, that provides

- a competence-centred, multi-source formative assessment methodology,
- based on sound psycho-pedagogical models (i.e., Competence-based Knowledge Space Theory and Formal Concept Analysis),
- intelligent model-based reasoning services,
- innovative visualization techniques,
- and features to open and negotiate learner models;

LEA’s BOX is dedicated to develop a learning analytics toolbox that is intended to enable educators to perform competence-centered, multi-source learning analytics, considering their real practical needs. Thus, the project spends significant efforts on a close and intensive interaction with educators in form of design focus groups and piloting studies.

The tangible result of LEA’s BOX manifest in form of a Web platform (Figure 1) for teachers and learners provide links to the existing components and interfaces to a broad range of educational data sources. Teachers will be able to link the various tools and methods that they are already using in their daily practice and that provide software APIs (e.g., Moodle courses, electronic tests, Google Docs, etc.) in one central location. More importantly, the platform hosts the newly developed LA/EDM services, empowering educators to conduct competence-based analysis of rich data sets. A key focus of the platform will enable teachers not only to combine existing bits of data but to allow

Figure 1. Central web platform.
them to "generate" and collect data in very simple forms, not requiring sophisticated hard- or software solutions. Finally, we want to open new ways to display the results of learning analytics - leaving the rather statistical dashboard approach, moving towards structural visualizations and towards opening the internal learner models.

3. Open Learner Models
Learner models contain and dynamically update information regarding users' learning: current knowledge, competencies, misconceptions, goals, affective states, etc. There is an increasing trend towards opening the learner model to the user (learner, teacher or other stakeholders) to support reflection, encourage greater learner responsibility for their learning, and help teachers to better understand their students [2]. The core requirement is that such visualizations must be understandable to the user. Although this may appear to be similar to the more recent work on LA, open learner models (OLM) concentrate more on the current state of learners, with less references to activities undertaken, scores obtained, materials used, contributions made, etc. OLMs typically focus on concepts, competencies, and guiding learners with regard to conceptual issues rather than specific activities and performance. Various OLM visualization examples have been described in the literature for university students (see [2] for a more detailed overview). The most common visualizations used in courses include skill meters, concept maps and hierarchical tree structures.

In addition to visualizing the learner model, various methods of interacting with the learner model exist, ranging from simple inspectable models, which allow some kind of additional evidence to be input directly by users, to negotiated learner models, in which the content of the learner model is discussed and potentially updated. We focus on the latter. Key features of negotiated learner models are not only that the presentation of the learner model must be understandable by the user, but also that the aim of the interactive learner modelling should be an agreed model. Most negotiated learner models are negotiated between the student and the teaching system. However, other stakeholders can also be involved, and the notion of "the system" can be broadened to include a range of technologies, such as the ones used in technology-enhanced learning. Here we consider (i) fully-negotiated learner models; (ii) partially-negotiated learner models; and (iii) other types of learner model discussion. They are all relevant to our notion of negotiating the learner model or its content, and they are adapted for LEA's BOX (Figure 2).

4. COMPETENCE SPACES
In the context of formative LA, a competence-oriented approach is necessary. Thus, a Hasse diagram can be used to identify and display the latent competencies of a learner in the form of so-called competence states. An elaborated theoretical approach to do so is Competence-based Knowledge Space Theory (CbKST). The approach originates from Jean-Paul Doignon and Jean-Claude Falmagne [6, 7] and is a mathematical psychological, set-theoretic framework for addressing the relations among problems (e.g., test items). It provides a basis for structuring a domain of knowledge and for representing the knowledge based on prerequisite relations. While the original Knowledge Space Theory focuses only on performance (the behavior; for example, solving a test item), its extension CbKST [1] introduces a separation of observable performance and latent, unobservable competencies, which determine the performance [1]. This is a psychological learning-theoretical approach, which highlights that competencies (e.g., the ability to add two integers) are unobservable latent constructs and which can only be observed or assessed indirectly.

We interpret the performance of a learner (e.g., mastering an addition task) in terms of holding or not holding the respective competency. In addition, recent developments of the approach are based on a probabilistic view of having or lacking certain competencies. In our example, mastering one specific addition task allows the conclusion that the person is able to add two numbers (to hold this competency) only to a certain degree or probability. When thinking of a multiple-choice item with two alternatives, as another example, mastering this item allows only to 50 percent that the person has the required competencies/knowledge.

On the basis of these fundamental views, CbKST is looking for the involved entities of aptitude (the competencies) and a natural structure, a natural course of learning in a given domain. For example, it is reasonable to start with the basics (e.g., the competency to add numbers) and increasingly advance in the learning domain (to subtraction, multiplication, division, etc.). As indicated above, this natural course is not necessary linear, which bears significant advantages over other learning and test theories.

As a result we have a set of competencies in a domain and potential relationships between them. In terms of learning, the relationships define the course of learning and thus which competencies are learned before others. In CbKST such relationships are called prerequisite relations or precedence.
5. VISUALIZING COMPETENCE SPACES

Hasse diagrams are capable of holding a number of important information for an educator to evaluate the learning progress and also to make recommendations. In this paper we want to highlight such advantages.

5.1 Competence States and Levels

As outlined, a competency space is the collection of meaningful states a learner can be in. Depending on the domain, the amount of possible states might be huge. The big advantage, however, is that depending on the degree of structure in the domain, by far not all possible combinations of competencies are reasonable and thus part of the space. When zooming into the diagram, a teacher can exactly identify the set of competencies that is most likely for the learner, by zooming out color-coding can illustrate the most likely locations of a learner within the space. When looking at the entire space, it is obvious at first site at which completion level a learner is approximately (rather at the beginning or almost finished). These zoom levels are shown in Figure 4. Technically, there is a variety of options to achieve the coding, for example, bolding, greying, or color coding, whereas likely states are displayed more distinctly than such with low probability.

Equal to individual states, Hasse diagrams can represent group distributions. Defined by a certain confidence interval of probabilities those states and areas can be made more salient that hold the highest percentage of learners of a group. By this means, specific areas in the competency space become apparent within which the most learners are and, in contrast also positive or negative outliers pop out the diagram. A different method was suggested by [10], who altered the size of the nodes to represent the groups’ sizes; the larger a node the more learners hold a particular state.

5.2 Learning Paths

In addition to having insight into groups’ and individuals’ current states of learning, the learning history, the so-called learning paths, are of interested for educators; on the one hand for planning future activities, on the other hand, for negotiation and documenting the achievements of a learning episode (e.g., a semester). Learning paths can be simply displayed by highlighting the edges between the most likely state(s) over time. As for the states, various probable paths can be realized by making more likely paths more intensive (by color coding or line thickness). Figure 5 shows a simple example. A key strength of presenting learning paths, as indicated, is opening up the learner model to the learners (perhaps parents) themselves [10] – to explain where they started at the beginning of a course and how they proceeded during the course and which competencies they hold today. This perhaps can be complemented with comparisons to others or groups. Not least, learning paths can unveil information about the effectiveness and impact of certain learning activities, materials, or the teacher herself.

6. CLASSROOM DATA COLLECTION

The features of Hasse diagrams and the arising advantages for LA appear all well and good. However, the key question is, where do they data for computing the probabilities of competence states come from. And everything stands or falls with this question. As for all techniques of LA, it depends on a data rich approach to education, the more and the better data exist, the better is the quality of LA conclusions. CbKST and Hasse diagrams are no exception to that. However, the approach of separating latent competencies, which more or less develop and exist in the black box ‘human brain’, and the performance they determine, bears particular advantages. On the one hand, performance, e.g. test scores, classroom participation, homework, etc., is not only determined by competencies or aptitude; there is a variety of aspects contributing to a certain performance, e.g., motivation, daily constitution, tiredness, external distractors, nutrition, health status, etc. On the other hand, CbKST-ish competence spaces are rather stable, once set up and validated properly. The advantage lays in the fact that performance such as test results, behaviors, achievements, etc. is considered as probability-based indicators for certain competencies. Mathematically this relationship is
established in form of interpretation and representation functions [1], which links an arbitrary set of performances/behaviors to one or more competencies, either in an increasing or in a decreasing sense. This, in the end, allows linking all available and perhaps changing data sources to one and the same competence space. It’s not about a single test, it’s about all available information we can gather, even it is considered being of little importance, all sorts of information may contribute to strengthen the model, the view of the learner. In case the amount or quality of data is weak, CbKST allows conservative interpretations, based on the arising probability distributions, in case there is a richer data basis, the probability distributions are more reliable, valid, and robust. For the educator, and this is important, the uncertainty is mirrored in the degree of likelihood. On a weak data basis, the probabilities of competence states differ substantially less than on the basis of richer data. Such information, however, can change the educator’s view and evaluation of a student’s achievements. In the end, this approach supports a fairer and more substantiated approach to grading or providing formatively inspired feedback.

7. CONCLUSIONS AND OUTLOOK
There is little doubt that frameworks, techniques, and tools for LA will increasingly be part of a teacher’s professional life in the near future. The benefits are convincing – using the (partly massive) amount of available data from the students in a smart, automated, and effective way, supported by intelligent systems in order to have all the relevant information available just in time and at first sight. The ultimate goal is to formatively evaluate individual achievements and competencies and provide the learners with the best possible individual support and teaching. Great. The idea of formative assessment and educational data mining is not new but the hype over recent years resulted in scientific sound and robust approaches becoming available, and usable software products appeared. However, when surveying the educational landscape, at least that of the EU, the educational daily routines are different. We face technology-lean classrooms and schools, we face a lack of proper teacher education in using ICT in schools – not mentioning of using techniques of LA in schools. We face a certain aloofness to use breaking educational technologies and a well-founded pedagogical view that learning ideally is analogous and socially embedded and doesn’t occur in front of some kind of electronic device. These are all experiences and results of a large-scale European research project named Next-Tell (www.nexttell.eu) that was looking into educationally practices across Europe and that intended to support teachers where exactly they are today with suitable ICT as effective and as appropriately as possible.

The framework of CbKST offers a rigorously competence-based, probabilistic, and multi-source approach that accounts for the latent and holistic abilities of learners and therefore accounts for the recent conceptual change in Europe’s educational systems towards a more competence-oriented education including multi-subject competencies and superordinate 21st century (soft) skills. No matter if data are rich or lean, a teacher is supported to the best possible degree and with a variety of important information about individual and group-based learning processes and performances and not least about the performance of learners and about the educator’s own performance. The probabilistic dimension allows teachers to have a more cautious view of individual achievements – it might well be that a learner has a competency but fails in a test; vice versa, a student might luckily guess an answer.

From an application perspective, in the context of European projects we developed and evaluated tools that cover the techniques and approaches described in this paper. In the Next-Tell project, for example, we developed a software tool named ProNIFA, which allowed linking multiple sources of evidence of learning and building CbKST-based learner models. We piloted various school studies and gathered feedback from teachers. In the end, and this can be considered an outlook for future developments, we had to find out that the ‘massive’ Hasse diagrams are overburdening teachers’ understanding and mental models about individual and class-based learning. Moreover, in order to understand the classical Hasse diagrams, it required (too) massive efforts in training teachers to fully utilize the potentials of those diagrams. Large scale surveys yielded that most educators still prefer simple but information-wise shallow visualizations such as traffic lights or bar charts significantly over more information-rich approaches such as Hasse diagrams or, just to mention another interesting approach, parallel coordinates. Therefore, recent efforts, e.g., in the LEA’s BOX (www.leas-box.eu) project, seek to adjust and advance the classical Hasse diagrams to such visualizations that are intuitively understood by educators and, at the same time, hold the same density of information. In particular, focus of research is on an advancement of Hasse diagrams towards specific mental models teachers may hold, such as a starry night sky or organic, biological structures such as cells of a living being. Also, abstraction and simplification techniques are investigated, e.g., fisheye lenses or streamgraphs. In conclusion, the utility of CbKST-ish approaches to LA, involving a separation of latent competencies and observable behaviors/performance, as well as having a conservative, probabilistic, multi-source approach appears to be a striking classroom-oriented, next-level contribution to LA, learner modelling, and model negotiations.

8. ACKNOWLEDGMENTS
This work is based on the Next-Tell project and the LEA’s BOX project, contracted under number 619762, of the 7th Framework Programme. This document does not represent the opinion of the EC and the EC is not responsible for any use that might be made of its content.

9. REFERENCES
Detecting Cheaters in MOOCs Using Item Response Theory and Learning Analytics

Giora Alexandron  
Massachusetts Institute of Technology  
giora@mit.edu

Zhongzhou Chen  
Massachusetts Institute of Technology  
zchen22@mit.edu

ABSTRACT
The focus of this work is on developing a general method for identifying cheaters in MOOCs in a way that does not assume a particular method of cheating. For that, we develop a classification model that takes as input a set of features that operationalize performance and behavioral parameters that are known to be associated with cheating. These include students’ ability, the level of interaction with the course resources, solving time, and Item Response Theory (IRT) person fit parameters. We start with a list of six candidate features, and after a feature selection process, remain with four. We use these to build a probabilistic classifier (logistic regression) that yields an Area Under the Curve (AUC) of 0.826. Our data is based on an Introductory Physics MOOC. The features are computed using data-mining and standard IRT packages. We consider only the users who received a certificate in the course. Each of these users is considered as an example for the classifier. The positive examples are the set of users who were detected as “using multiple accounts to harvest solutions” by a different algorithm that was reported in a previous publication.

CCS Concepts
●Applied computing → Interactive learning environments; E-learning; ●Information systems → Data mining;

Keywords
Academic dishonesty; MOOCs; learning analytics; Item Response Theory

1. INTRODUCTION
Academic dishonesty is a serious problem, with studies reporting that up to 95% of college students are engaged in academic dishonesty of some form [3, 8, 9, 15, 18]. In online setting, Palazzo et al. [16] found that between 3 and 11% of the submissions in an interactive online learning system were copied.

Massive Open Online Courses (MOOCs) are a relatively new domain, with certificates that currently do not have formal value (except for few pilot programs). However, several studies already reported the non-surprising findings that cheating exists also in MOOCs. According to [1, 2, 14, 17], between 1 to 10% of the students are using multiple accounts to harvest solutions. This cheating method was dubbed CAMEO (Copying Answers using Multiple Existence Online [14]; We refer to a person who uses this method as CAMEO user).

The amount of cheating that has been reported so far for MOOCs involves only the use of CAMEO and requires that the master and the harvester accounts can be linked by IP. Since there are certainly other forms of cheating in MOOCs (including performing CAMEO using accounts not linked by IP), the above is only a lower bound to the actual size of this phenomenon.

The main risk posed by cheating is decreasing the perceived value of the MOOC certificates, since a significant amount of cheating reduces the confidence that the certificate truly reflects students’ ability. For example, we found that students who used CAMEO gained almost half of standard deviation in their IRT ability by using this method.

Another risk of cheating is affecting the results of educational research [1, 17]. We found that CAMEO users had better performance, both in terms of success rate and response time, than the rest of the certificate earners in the course. In addition, the CAMEO users that we observed tended to do a lot of questions, but not to interact a lot with the instructional materials. This might lead to a false conclusion that in our course it is better (or even suffice) to spend time on doing questions, rather than learning from the instructional materials.

MOOC providers acknowledge the fact that cheating is a problem that they need to address, and use various proctoring systems. Currently these systems are mainly designed against impersonating. They are not effective, for example, against CAMEO, and probably also against other methods that are still unknown.

The goal of this work is to bypass the possibility of students designing methods to specifically thwart the CAMEO detectors by developing a general detection method that is not tailored to a specific form of cheating, but rather relies on measuring aspects of behavior that are either associated with or affected by cheating. The aspects that we currently consider include the amount of interaction with the course resources, time to answer, student’s ability, and two person-fit parameters obtained from IRT – Guttman error [10], and the standard error of ability estimates.

The rationale for using the amount of interaction with the resources is based on the assumption that cheaters will have less interaction with the instructional resources, as they do not need them to solve the questions on which they cheat.
Very fast time to answer was identified by [16] as a strong signal for cheating.

The rationale for using student’s ability is that cheaters tend to have a relatively high performance comparing to the rest of the certificate earners [17].

The rationale for using person-fit parameters is based on the assumption that cheaters have a relatively ‘noisy’ performance, as their performance depends not only on their ability, but also on whether they cheat or not. Following this rationale, researchers in the psychometrics community developed various person-fit indexes to measure unusual response patterns, including cheating [6, 11]. Among them, Guttman error, which measures the number of item pairs in which an easier item is answered incorrectly and a more difficult item is answered correctly, was shown by Meijer [10] to be a simple and effective person-fit index for identifying cheating. Thus, we use this parameter.

In addition, the standard errors of ability estimates in IRT model could also be used as a measure of unusual response patterns. The rationale behind using this measure is that an aberrant response provides inconsistent psychometric information, and thus leads to an increase in the standard error of the ability estimates [7].

Using these parameters, we train a probabilistic classifier (logistic regression) on data that contain 10% cheaters who used CAMEO, and were identified by algorithms that their description and verification process are described in detail in [1, 17]. On this data, the classifier achieves an AUC of 0.826.

To the best of our knowledge, this is the first study that suggests a general method for detecting cheating in MOOCs. It does so by combining machine learning, psychometrics, and learning analytics. Thus, we believe that the results are of interest for the educational data science research community, though these results are still preliminary.

The rest of this paper is arranged as follows. In Section 2 we present in detail the data and the method. In Section 3 we present the results. Discussion, limitations and future work are presented in Section 4.

2. DATA AND METHODS

2.1 Data

We use the data from the Introductory Physics MOOC 8MReVx given by the third and fourth listed authors in summer 2014 through edX.org. The course covers the standard topics of a college introductory mechanics course. It contains 273 e-text pages, 69 videos, and about 1000 problems (checkpoints problems embedded within the e-text and videos, and homework and quiz questions which are given at the end of the units). About 13500 students registered to the course, and from them, 502 earned a certificate. For this research, we considered 495 out of the 502 certificate earners (7 were omitted due to technical reasons). Among these 495 certificate earners, 65 were detected as CAMEO users (namely, users who harvested answers using multiple accounts) by the algorithm reported in [1], which is a modification of the algorithm presented in [17]. Both algorithms were verified using manual and statistical inspection methods (a full description of the algorithms and the verification process can be found in [1, 17]).

2.2 Feature selection

2.2.1 Predictors

We start with an initial set of predictors that divides into two groups:

**Behavioral parameters:**

i. Fraction of videos watched.

ii. Fraction of correct answers that were submitted in less than 30 seconds (the cutoff considered by [16]).

iii. Mean time for submitting a correct answer. (For ii and iii, the submission time is operationalized as the gap between the time of entering the page in which the problem resides, and the time the correct answer is submitted.)

The rationale for using these parameters is described in Section 1. These parameters were mined from the logs using standard scripts.

**Ability and person-fit parameters:**

iv. Student’s ability, computed by a two parameter logistic (2PL) model in IRT using the BILOG software package. The input to the IRT algorithm is a binary response matrix computed from students logs. The response matrix contained only the certificated students (accounts), and items that were answered by at least 50% of these students.

v. Guttman error – the number of item pairs in which an easier item is answered incorrectly and a more difficult item is answered correctly.

vi. Standard error of student’s ability estimate from IRT.

The two person-fit parameters – Guttman and standard error, where computed using the output of the 2PL model used to estimate students’ ability.

2.2.2 Dependent variable

It is a binary variable that indicates whether the student is a CAMEO user, namely, used multiple accounts for harvesting solution in our course. The positive examples are the accounts that were identified as CAMEO users by the algorithms described in [1, 17].

2.2.3 Initial feature set

For each user, we build an example vector containing the values computed for this student for parameters i-vi, and the cheater/non-cheater tag. Together these form the feature set.

2.2.4 Standardizing the data

The independent variables were standardized using z-scores, so that we can compare the relative importance of features based on standardized logistic regression coefficients [12].

2.2.5 Removing redundant features

To remove redundant features, we use a L1 regularized logistic regression and pick the features that have a non-zero coefficient [5]. This is implemented using R’s glmnet package [4]. The features that are found to be redundant are the mean-submission-time (iii), and the ability parameter (iv).

2.2.6 Final feature set

After removing the redundant features, we remain with a feature set containing four predictors: Standard error for IRT student ability parameter, Guttman error, fraction of videos watched, and fraction of questions answered in less than 30 seconds.
2.3 Classification model

We use this set to build probabilistic classifier using a logistic regression. The classifier is evaluated by examining the area under the ROC curve, using k-fold cross-validation. The results are presented below.

3. RESULTS

Below we present, per feature, the difference in the distribution of the values among cheaters and non-cheaters, and the results of the classifier that is built on these features.

3.1 Difference between cheaters and non-cheaters – individual parameters

For each of the four features, there is a statistically significant difference between the cheaters and the non-cheaters (p-value < 0.001 for all the features). Table 1 shows, per feature, the standardized mean value for each group.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cheaters</th>
<th>Non-cheaters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error</td>
<td>0.36</td>
<td>-0.05</td>
</tr>
<tr>
<td>Guttman error</td>
<td>0.94</td>
<td>-0.14</td>
</tr>
<tr>
<td>Very quick answers</td>
<td>0.99</td>
<td>-0.14</td>
</tr>
<tr>
<td>Videos watched</td>
<td>-0.77</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The distribution of these values is presented in Figure 1. The upper left figure shows the fraction of questions that were solved (correctly) in less than 30 seconds. For non-cheaters, this percentage is very small (the blue sharp curve), while for cheaters, this is much higher. The right upper figure shows the distribution of the fraction of videos watched. The red sharp curve shows that most of the cheaters watched a very small fraction of the videos, relative to the non-cheaters. The bottom left curve shows the standard error for the ability parameter. As can be seen, non-cheaters tend to have a smaller error. However, for this feature the distinction is less clear. The bottom right figure shows the distribution of the Guttman error. Again, non-cheaters tend to have lower Guttman error.

In all the figures, the relative smoothness of the cheaters’ curve might be related to the differences in the amount of cheating among the cheating students (ranging from 1% to more than 50% of the correct answers).

3.2 Performance of the classifier

We used the feature set to build a logistic regression classifier, using R’s e1071 package [13]. The performance of the classifier was evaluated by examining the area under the ROC curve (AUC), using a 3-fold cross validation. We pick k = 3 because the data is biased (about 10% cheaters and 90% non-cheaters), and we want to ensure that in each iteration, with high probability both the training and the test sets will include sufficient number of positive examples (cheaters).

**AUC.** Overall, the AUC of the model was 0.826, with a variance of 0.0016. These results are for 3-fold cross-validation, ran for 500 times.

**ROC curve.** Next, we divide the data at random to 2/3 training set and 1/3 test set. The AUC on the test set of a model built on the training set was 0.852. Figure 2 shows the ROC curve for this model.

To compute the optimal cutoff, we look for the optimal point on two metrics. One is the cutoff that minimizes the distance from the ROC curve to the optimal classification point (0,1). The other is the cutoff the maximizes the sum of the true-negative and the true-positive rates. The cutoffs are 0.148 and 0.149, respectively.

4. DISCUSSION

Our results show that a probabilistic classifier that uses four features – two IRT person fit parameters (Guttman error and standard error), and two simple learning analytics parameters (fraction of videos watched and fraction of correct answer in less than 30 seconds), can detect users who used multiple accounts to collect correct answers (‘CAMEO users’) in the 2014 run of 8.MReVx MOOC with a good level of accuracy (AUC of 0.826).

The crux of our approach is using ‘circumstantial evidence’ that is associated with cheating, but is not specific to a certain method (e.g. CAMEO). Thus, we believe that such a model can identify students who use other forms of cheating. One of the steps that we take to evaluate the feasibility of this approach is examining the logs of the ‘false positives’ – accounts that are identified by the algorithm as...
cheaters, but were not detected as CAMERO users by the CAMERO algorithm.

Our analysis indicates that at least some of these ‘false positives’ are students who are identified by the CAMERO algorithm as ‘suspicious users’ but are filtered because they also behave as harvesters (the accounts that are used to collect the correct answers). We believe that these are actually users who collaborate with each other, and for example divide some of the work between them (and thus their account sometimes appears as the account that collects the answers, and sometimes as the accounts that uses the answers collected by another account).

We regard it as likely that our methods will generalize to other courses, as we see no reason to believe that the Introductory Physics course that we studied is specifically attractive to cheaters. Thus we expect to see cheating in other MOOCs as well, and we believe that this will be also associated with performance and behavioral patterns that could be used to distinguish between cheaters and non-cheaters.

4.1 Limitations

Generalizing the results. The main limitation for generalizing our results to other courses is the fact that our data is based on one course. Generalizing to (identifying) other methods of cheating is limited by the fact that we trained our model on data that includes cheaters who used a specific method.

Post-factum analysis. The approach that we present in this paper is post-factum in nature, and is less suitable for identifying cheating events as they occur. Because it relies on ‘circumstantial evidence’, rather than on a direct evidence for a specific kind of cheating, it is required to accumulate a considerable amount of evidences in order to achieve a sufficient level of confidence. However, it seems reasonable to assume that this can be done during the course (for example, at the end of each chapter).

4.2 Future work

Main directions for future research include studying additional features (e.g., person-fit and behavioral parameters) that can be used to improve the classification, and extending the study to more courses and other forms of cheating.

5. ACKNOWLEDGMENTS

The authors would like to thank Jose Antonio Ruiperez-Valiente for his help in conducting this research.

6. REFERENCES


WiBAF into a CMS: Personalization in Learning Environments Made Easy

Alejandro Montes García
Eindhoven University of Technology
Eindhoven, the Netherlands
a.montes.garcia@tue.nl

Paul De Bra
Eindhoven University of Technology
Eindhoven, the Netherlands
p.m.e.d.bra@tue.nl

Natalia Stash
Eindhoven University of Technology
Eindhoven, the Netherlands
n.v.stash@tue.nl

George H. L. Fletcher
Eindhoven University of Technology
Eindhoven, the Netherlands
g.h.l.fletcher@tue.nl

Marc Fabri
Leeds Beckett University
Leeds, United Kingdom
m.fabri@leedsbeckett.ac.uk

Mykola Pechenizkiy
Eindhoven University of Technology
Eindhoven, the Netherlands
m.pechenizkiy@tue.nl

ABSTRACT

Adaptivity has proven successful in reducing navigation and comprehension problems in hypermedia documents. Authoring of adaptive hypermedia documents and especially of the adaptivity in these documents has been problematic or at least labour intensive throughout AH history. This paper shows how the integration of a CMS with an adaptive framework greatly simplifies the inclusion of personalization in existing educational applications. It does this within the context of European project Autism&Uni that uses adaptive hypermedia to offer information for students transitioning from high school to university, especially to cater for students on the autism spectrum as well as for non-autistic students. The use of our Within Browser adaptation framework (WiBAF) reduces privacy concerns because the user model is stored on the end-user’s machine, and eliminates performance issues that currently prevent the adoption of adaptivity in MOOC platforms by having the adaptation performed on the end-user’s machine as well (within the browser).

Authoring of adaptive applications within the educational domain with the system proposed was tried out with first year students from the Design-Based Learning Hypermedia course at the Eindhoven University of Technology (TU/e) to gather feedback on the problems they faced with the platform.

CCS Concepts

• Information systems → Web applications; • Software and its engineering → Software creation and management;

Keywords

adaptation, learning styles, content management systems, software development

1. INTRODUCTION

Adaptive hypermedia has been successful for many years in improving the ease of navigation and the user comprehension of the content on Hypermedia documents [3, 5, 7]. However, adaptive hypermedia needs more effort from the development/authoring team than hypermedia applications without adaptive features [4]. While this is sometimes seen as a burden, the effort that would be needed to have equally good navigation and content comprehension properties in a non-adaptive hyperdocument would be far greater. The scientific community has been trying to minimize the extra development and authoring effort for years by using custom-made languages [11, 13, 20] or graphical interfaces [19]. Even without considering adaptation, relatively few people can develop HTML applications directly through writing in HTML syntax and even fewer people can add adaptive behaviour to them.

Content Management Systems (CMSs hereinafter), have become very popular tools for the development and maintenance of web applications, mainly because they enable people without any knowledge of HTML, CSS or other (client- or server-side) web languages to design professional looking websites. Some of these CMSs are free and open source, allowing people to develop all kinds of plugins or modify the core of the CMS to extend its functionality. According to a survey made by the website http://w3techs.com/, as of April 5th, 2016, 44.4% of the web pages that they track run one of no less than 313 CMSs that they monitor. Among those WordPress1 is the most popular one, present in 26.3% of all the sites tracked.

CMSs have helped many people to build their own website without in-depth knowledge of web technologies, yet there are no CMSs that take effective advantage of adaptive hypermedia extensions. Existing Adaptive Hypermedia Systems may try to make the creation of adaptive websites feasible, but no current AHS achieves the user-friendliness of popular CMSs. Hence we turn the process around: from creating adaptive websites, we move to making websites adaptive. To do so an existing CMS needs to be extended with adaptive functionality. This paper describes such an extension: integrating adaptive functionality in WordPress. It also describes the use of this functionality in the European Autism&Uni project2 to adapt information to students on the autism spectrum, specifically for helping them with the transition from high school to university. The adaptive functionality we used is more generic, allowing for adaptation to different learning styles and to general “prerequisite” relationships between topics. The WiBAF (Within-Browser Adaptation Framework) extension to WordPress also takes care of typical privacy concerns of users by performing user modelling and adaptation entirely within the browser (and thus on the end-user’s device rather than on a server).

In order to develop our work we have taken the following decisions:

• We used WordPress as our CMS because it is free, open-source and the most used in real-life environments.

1 see https://wordpress.org/
2 see http://www.autism-uni.org/
• We used WiB AF [12] as our adaptation library because the client and server code are decoupled and it is easy to modify the server code to make it compatible with WordPress. (WiB AF is not limited to be used with WordPress and has been also used to add adaptation to websites created with different CMSs, or without any CMS.)

• We focused on an on-line educational setting. In order to perform adaptation in such setting, we focused on how the student learns, what the student has already learned and what is still to come.

With these decisions taken, our goal became to integrate WiB AF and WordPress in such a way that personalization can be easily added to existing educational sites created using WordPress. At the same time the adaptation should be done efficiently and it should track how the user learns and what she has already learned. The efficiency needed for massive applications like MOOCs (Massive Open Online Courses) can actually be obtained because almost all of the work is done in the browser, thus avoiding server-side performance bottlenecks.

Therefore, our contribution with this paper is the integration of an adaptation library in a Content Management System in order to ease the development of adaptive educational applications. This can be applied to current existing applications like MOOCs in an easy way. This means that rather than developing applications from scratch, our tool allows to make existing applications adaptive.

The remainder of this paper is structured as follows: We describe our use case scenario in Section 2 and then we describe how the integration between WiB AF and WordPress has been made in Section 3. Section 4 explains how to develop adaptive educational applications, both for existing applications and new ones. Later, in Section 5, we describe the experiences of first year students with the tool. The related work is presented at Section 6. Finally, we conclude and propose future work in Section 7.

2. USE CASE SCENARIO

Before describing how the integration has been done, we introduce our use case scenario to make the requirements we have come up with more concrete. It is worth mentioning here that the platform developed is generic and as such can be applied in different domains. Recently we have used it in the educational context in our first year course Design-Based Learning Hypermedia, in particular for creating the “First Aid Kit” for students entering the university.

We set forth to use the WordPress and WiB AF integration in the Autism&Uni project. This project is aimed at widening access to higher education for autistic students by providing a toolkit that can help them overcome the challenges they may face when going to university. The goal is to give students a taste of how higher education works and how to cope with the physical university environment before they start their study. This toolkit1 is offered as an Adaptive Web-Based Application.

The adaptive functionality differentiates in how the information site presents itself to autistic and non-autistic students, but in the end the toolkit provides the same information to everyone.

Adaptation for autistic students is concerned with adapting to the differences in cognitive abilities, within this project in particular comprehension, between autistic and non-autistic people. In our previous work [14] we discussed that it seems like autistic people have their own ways of processing and analysing information, their own ways of processing, analysing information and learning and as such they can be considered as having a specific cognitive-

1http://www.autism-uni.org/toolkit/

or learning style. In order to provide effective adaptation, we utilize the specific characteristics and preferences of the user in three different learning styles dimensions [10], i.e. where is the user located in the: visual vs. verbal axis, global vs. analytical axis and active vs. reflective axis. We make use of the user history as well. These variables together with the adaptation effects provided by the toolkit, have been described in our previous work [8].

A secondary, but also important aspect of performing adaptation in the presence of autistic users is the heightened awareness of (and anxiety for) the user modelling involved. Autistic students do not only experience anxiety when entering an unknown environment, but also when they realize that their personal and possibly sensitive data are stored on an external computer that they cannot access, when they do not feel their data is kept private or they cannot control it. Fortunately WiB AF stores all user data on the client side (using browser storage) by default. Autistic users may choose to keep this setting, thus guarding their privacy, while other users may opt to share their data in order to enable the server side to perform group adaptation. (We currently do not offer group adaptation, but we do offer the user model sharing option).

For this specific use-case, we also consider some factors related to the context, namely where the student is and what time it is. The reason for this is that autistic students often feel lost, they need reminders that tell them where they have to go.

We are implementing a feature so that they can import events from their Google Calendar. The tool will show a reminder when the student needs to go to a lecture and a link with the instructions on how to get to the room where she needs to be. This is still under development and not yet part of the generic platform, therefore we will not describe it further. We mention the notification feature because it needs to be developed in order to really help autistic students.

In order to effectively display the content of our learning objects, we have broken it down into small pieces or fragments with some semantic meaning, from which the student can learn something. In our case, we show an introduction first, we show also a comic strip or an image that shows quotes of students about the topic of the learning object, establishing a context for it. Then some background information is provided to justify the learning object. After that we talk about how the learning object being described is important for the reader and what she should do. We close the article with some additional tips, questions to think about and some follow-on reading. Each learning object can also have an alternative video version as well as pre- and post-requisites.

3. INTEGRATION OF WIB AF AND WORDPRESS

In order for WiB AF to understand the content of a learning object, the “sections” of a page need to have a unique identifier within that page. WordPress allows developers to create custom themes to change the look and feel and the structure of a page. We used this functionality and created a custom template that creates the sections and assigns ids that WiB AF can understand.

In order to make content providers aware of this division, we needed to modify not only the page template itself, but also the interface used to enter page content. We used a plug-in called Advanced Custom Fields2 or ACF for that.

WordPress also allows developers to create plugins in order to extend its functionality. We have packed all the functionality required for WiB AF in a plugin. We will describe now the functionality present in our plugin. (This knowledge is not needed to

2http://www.advancedcustomfields.com/
understand the authoring process.) WordPress offers a settings interface that has been modified to add new fields related to the user model. The data contained in those fields are stored in the client by default. WiBAF offers users the possibility to send these data to the server if desired. That could make it possible to (potentially) create better personalization (through collaborative filtering) or to provide a facility to synchronize their data between devices. By default the data are stored on the browser only, to offer better privacy. We also needed to add these new fields to the user data so that they can be stored on the server (if desired). The management of the database is done automatically by WordPress, we just need to declare that these extra data exist. The communication between server and client to manage user data is done as follows: when (and if) the user wants to send data to the server, this is done through cookies. These cookies are created using JavaScript after the page is loaded and WiBAF has finished parsing and executing the modelling and adaptation files. Therefore, the cookies will contain the user values after visiting the page, but every interaction that the user might do that can change the user model, will not be sent in this cookie, but in the following one, after another request. When the server wants to update the client values (e.g. the user logs-in from a new device and she wants to keep her profile), it inserts the JavaScript code required to do so. This code is inserted just after the cookie creation and updates its values. This might not seem efficient, but because the values are stored asynchronously in the IndexedDB, it is more efficient to create the cookies and then update them when the database has finished the insertion operations.

The content of a website offered through a CMS is very dynamic because it is very easy to create and update. Therefore, we need to replicate this dynamism in our adaptation code. To do so, we have to update the adaptation automatically when a new learning object is created. We achieve this by using action hooks. The action hooks provided by WordPress are not only related to special actions. In every page request, typically around 50 actions are executed. Code can be hooked to a specific action. This code will then be executed when that action occurs. In our plugin we hooked code to several actions to link the WiBAF JavaScript library, the adaptation files, and to store user data in the browser (or on the server if the user wants that). It is worth mentioning that we also “taught” WordPress to write adaptation code for us. Every time a new learning object is created, WordPress writes some lines of adaptation code to track it.

4. DEVELOPMENT OF ADAPTIVE APPLICATIONS

As we mentioned before, the integration of WiBAF and WordPress was built so that creation of adaptive applications from scratch is possible, but also adding adaptive behaviour to an existing application is possible. This means that the tool is useful for creating adaptive applications, but also for making existing applications adaptive. While the tool is generic, our implementation was built having in mind the educational domain, therefore in the following subsections we will explain how to create new adaptive applications in the educational domain and how to make an existing educational application adaptive.

4.1 Creating new adaptive applications

In order to create an adaptive application from scratch in our WordPress, the workflow is quite similar to the process of creating a normal WordPress application as described in the official tutorial. However, because we have divided our content in small fragments

---

[1]https://make.wordpress.org/support/user-manual/content/

---

Figure 1: Custom fields as shown on the dashboard.

(sections), this needs to be considered both in the template and in the dashboard. As we mentioned before, we used the ACF plugin to achieve this in the dashboard and we modified the post template in our theme so that it is consistent with the ACF settings. Figure 1 shows the view from the dashboard of the fields; as it can be seen each field should be filled according to the instructions in order to create a learning object.

Once the learning objects are created, the authors can proceed to add the adaptive behaviour to the existing application, as described in the next subsection.

4.2 Adding adaptive behaviour to an existing application

In order to add adaptive behaviour to an existing application, WiBAF can be added as a WordPress plugin. This can be done as shown in the official WordPress tutorial.

Once the plugin is installed, WiBAF can be configured from the dashboard as shown in Figure 2. The privacy levels can be configured, as well as the path where the adaptation and modelling files are located. These files follow the format described in our previous work [13].

The adaptation and modelling files can be modified in order to fit other needs, but the files we provide automatically track the user learning styles as we defined in our requisites, as well as her user history and it can react to it by hiding the unsatisfied pre-requisites, or by re-ordering fragments of the learning object.

---


---
WiBAF also offers users control over where its user model is stored. This is done by using a slider with three or more levels. In the first level of the slider, no data are tracked and therefore no personalization is offered. In the second level, all the data is stored on the browser. In the last level, all the data is stored on the server. The WiBAF plugin makes it easy to add new levels between the second and the last, in a way that only certain types of fields are sent to the server, while others are kept private in the browser, always depending on the user preferences.

5. EVALUATION WITH STUDENTS

Our first year students from Design-Based Learning Hypermedia course have used a preliminary version of the WiBAF and WordPress integration in order to develop adaptive educational applications. The most substantial difference between the version that the students used and the one described here is that WiBAF was not offered as a plugin, but its code was closely tied to the WordPress theme used (by the Autism&Uni project). It is worth mentioning that while some of the students that took the course had learnt some programming by themselves, most of them did not develop any web application before, not to mention an adaptive educational web application. After finishing the course, we invited them to complete a survey about their experiences with the tool. In total, 36 students completed the survey. They made it obvious that a plugin needed to be created, because the most difficult task for them was to extract the changes that made WordPress adaptive from the Autism&Uni theme and incorporate them into their own theme. 69% of the students found this task “hard” or “very hard”. Figure 3 shows the distribution of the opinions of the students regarding this task. With a plugin, this should become easier. Once the students managed to have the whole environment setup, and the WordPress customized as they wanted, the percentage of students that found adding adaptation “hard” or “very hard” was reduced from 69% to 22%, with 28% of the students finding it “easy” or “very easy” while the rest judged the difficulty of developing an adaptive educational application as neither hard nor easy. This can be seen in Figure 4. A similar result was obtained in another similar evaluation of GALE [6], in which the installation aspects were considered harder than the development of applications itself.

We also asked students to estimate the extra time that they had to spend in order to make the application adaptive, as compared to the...
development of a normal application. 48% of them thought that enabling personalization in an educational application takes 20% more time than a non-adaptive educational application while 69% of them estimated that the overhead introduced during the development was of about 50% or less. We think that these numbers will be improved even further, because as we mentioned before, the harder task has been greatly simplified. However, these numbers may not be very significant because the first year students had never before seriously considered adaptation, so more “serious” adaptation authoring may take more work because it requires more careful design by the authors.

6. RELATED WORK
Adaptive Hypermedia is a research field that can be traced back to the nineties [2]. It has become more complex since then and several new frameworks have been developed. They aim to ease the development of this kind of applications. Some good examples of those frameworks are AHA! [9] or GALE [17]. The work presented here takes an approach more similar to InterBook [4] in the sense that authoring is done using a standard tool (that was Microsoft Word for InterBook) and adding the “hints” for adaptation based on the student’s knowledge and on prerequisites was only a small extra effort. We also take into account student knowledge by checking the browsing history, but we extend that with adaptation to cognitive/learning styles and further we combine it with the simplicity of CMSs. We also provide tools to give control to the users about where their data is stored, either in the client or in the server, with all its advantages and disadvantages. Another authoring tool related to our work is the Dynamic Courseware Generator [21], which allows to create courses that adapt to the student knowledge. It also allows students to define their own learning goals and the tool will recommend a path to follow. One more research effort worth mentioning here is ALEF [16]. ALEF divides each learning object into fragments and uses strategies to re-order content; it also supports different types of learning objects such as explanations, questions and exercises and adds metadata to the content to produce the resources that are consumed by the students. However, it does not deal with the authoring of the adaptation itself.

Cognitive/learning styles refer to the different ways a person processes, analyses information and learns. There is previous research on adaptation to cognitive/learning styles and how these can be incorporated into Adaptive Hypermedia Systems and e-learning platforms [15, 18]. While adaptation to learning styles is useful in every e-learning platform, this is especially important in our use case scenario with autistic students as we showed in our previous work [8, 14].

To the best of our knowledge, little work has been done about bridging Content Management Systems and Adaptive Hypermedia, at least when considering Adaptive Hypermedia to be a broader concept than responsive design (which is adaptive only to the user’s browsing platform). The only work similar to what we have built was done by a master student in Eindhoven: Sander Brouwer [1]. In his work, a CMS with built-in adaptive characteristics was developed from scratch. This makes it more efficient than our approach, but it is unlikely that real life production environments would adopt a new CMS that does not have a big community solving issues or developing plugins, like WordPress or Joomla.7 This work also required authors to write the adaptation rules themselves while in our work, the framework that we developed is already adaptive to the user history and learning styles and has a straightforward usage in the educational domain.

An important new development in learning is the publicly available MOOCs. The massive nature of MOOCs requires high performance platforms for serving content, perhaps even using server-side page caching (to eliminate most of the page regeneration time). Most common adaptation platforms require that the adaptation is performed on the server side, resulting in different page content for each user, and different presentation of links on the pages. By using WiBAF the user modelling and adaptation effort can all be performed inside the end-user’s browser, thereby significantly lowering the computation to be done on the server side. As a result, the approach taken by WiBAF is the only feasible solution for adding adaptation to MOOCs. There may be (still unreported) ongoing work on adaptation for MOOCs but no on-line courses that are really massive are currently adaptive.

7. CONCLUSIONS AND FUTURE WORK
In this work we have shown how personalization can be enabled in existing Content Management Systems in order to ease the development of Adaptive Web Based applications in the educational domain. We have limited our research to the adaptation in e-learning environments, but this could be extended to different domains. We automatically track the user knowledge via her browsing history as well as user learning styles, looking at what users do with the content provided. Then we adapt the content using those user features. By integrating WiBAF into WordPress, we also allow developers to make existing applications adaptive, besides making adaptive applications, thus allowing content providers in focusing on the content rather than in the adaptation.

We tried our framework in the use case scenario for the Autism&Uni project and in the first year Design-Based Learning Hypermedia course at TU/e for creating the “First Aid Kit” for students entering the university. After following the course the students were asked to provide us with some feedback about their experience with the framework. This feedback has been taken into account to greatly ease the most difficult task pointed by the students.

Our framework allows for the development of adaptive applications by using learning styles and user history, but extra functionality can be supported as long as they are generic for educational applications. This functionality should be added to the plugin code. Non-generic functionality can also be implemented by editing the adaptation and modelling files used in WiBAF.

8. ACKNOWLEDGMENTS
This research is supported by the project WiBAF and the Life Long Learning Programme (project no. 10018300 Authew 539031-LLP-2013-1-UK). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any user, which may be made of the information contained therein. The authors thank the Autism&Uni project team, TU/e psychologists specialising in autism, TU/e students with autism who participated in interviews with us and toolkit evaluation, Stichting Handicap+Studie9 and Autistic10 for their valuable input.

9. REFERENCES


Towards Personal infrastructure to manage long term open learner models

Jean-Marie Gilliot
Institut Mines-Telecom, Telecom-Bretagne, UMR CNRS 6285 LabSTICC
jm.gilliot@telecom-bretagne.eu

Nour El Mawas
Institut Mines-Telecom, Telecom-Bretagne, UMR CNRS 6285 LabSTICC
nour.elmawas@telecom-bretagne.eu

Serge Garlatti
Institut Mines-Telecom, Telecom-Bretagne, UMR CNRS 6285 LabSTICC
serge.garlatti@telecom-bretagne.eu

ABSTRACT
Nowadays, the lifelong learning is a key issue in our life. Learners have their personal learning data scattered on different platforms and websites without any control on them and without any defined access duration. In this paper, we propose to explore the feasibility of Personal information manager systems in the Open Learner Model context that allows the control of personal learning data by learners themselves, the persistence, and the privacy. We propose to focus on a relevant technical infrastructure giving full personal control to users in order to manage Open Learner Models in lifelong and life wide perspectives. This work is dedicated for all lifelong learners without any specific IT competency to manage their own personal learning data in a lifelong perspective.

Keywords
Open Learner Models; Personal Cloud; Personal data; E-portfolio; Lifelong Learning.

1. INTRODUCTION
We all know that a person will have many different jobs during his/her life. Lifelong Learning is becoming a central asset, beginning during initial training at university, pursuing during the whole career with many different jobs. Learning is also life wide as we recognize that it occurs in multiple contexts: school, home, work, etc. Lifelong and Life wide Learning are seen as key elements for the prosperity, especially in a knowledge society. In this context, learners’ personalization and social learning are essential concepts [1]. They encompass formal and informal learning in everyday situations, as well as lifelong goals management.

Learner Models are the representation of knowledge and learning process and they are also part of advanced learning environments. Open Learner Models (or OLMs) are Learner Models that allow the user (learner, teacher, peers and/or other stakeholders in the education process) to view the content in human-understandable form. They can also be Independent, or external to the system, giving the opportunity to the user to monitor, understand, and plan future learning throughout life. These may support reuse of parts of the Learner Models by different applications [2]. We consider that Independent Open Learner Models must be considered as long term models to encourage reflection, facilitate monitoring of learning and cooperation in social contexts.

According to [3], e-Portfolios are a form of the Open Learner Model which is the learner driven. A portfolio is a meaningful documentation of a learning path, either for assessment or for formative purposes [4]. E-Portfolios are one of those tools that have been appeared in education since Internet usage becomes more widespread. E-Portfolios represent an advantage over traditional portfolios in terms of storage, access, management, interactivity, real-time functionality, and presentation method. Compared with paper-based portfolios, they also have the added value in terms of keeping records, connecting ideas, relating information, and publication [5]. Research studies [6] have shown the e-portfolio influence and impact on learning performance. The e-Portfolios have a significant effect on education, they enable the aggregation and disaggregation of student data [7], which can then be used in program evaluation and accreditation [8].

Consider the following example; Alice is an engineer having completed twelve years of primary / high school in Australia, and four years at a school engineering in Belgium, and a final year project in Germany. She acquired Open Badges and certificates online. She developed assessed professional skills at work in different positions. She is recommended by many professionals on her LinkedIn profile. She also monitors her involvement in programming communities. She needs to access to all those models in her personal space, enabling her to collect data about any knowledge / skill and visualize progress.

This scenario illustrates a long term user model that aggregates data from many different sources, and is used in different contexts. Learning achievements and outcomes must be collected across different contexts: formal and informal learning (from institutional Learning Management Systems to personal quantified-self devices [9]), across different countries, and must remain available lifelong under control of the learner. The learner needs to monitor his Learner Models, modify them when relevant, store them for further use, publish them, and share them with peers.

These needs may be ensured by Personal Cloud features. The Personal Cloud describes a user-centric model of Cloud computing where an individual's personal content and services are available anytime and anywhere, from whatever device they choose to access it. And in emerging economies, where people often share mobile devices, each individual would be able to log into their own Cloud from the shared device. According to Frank Gillet, an analyst with Forrester Research, the Personal Cloud and how it will shift individual computing "from being device-centric to information-centric”. He concludes that digital devices and services will combine to create the Personal Cloud, “an internal resource for organizing, preserving, sharing and orchestrating personal information and media.”
New solutions like Personal Information Managers (PIMs) \[10\] are cloud-based data managers that provide data persistence and privacy-by-design \[11\] infrastructure. Cloud-based enables reliability of the storage, and access from everywhere. Security is ensured by design. PIMs are open source based and provide the ability to monitor networks exchanges, ensuring that third parties services meet their commitments.

In this paper, we propose to explore the feasibility of Personal information manager systems in the Open Learner Model context that allows the user control, persistence, enabling privacy as well as self-defined sharing. We propose to focus on a relevant technical infrastructure giving full personal control to users in order to manage Open Learner Models in a lifelong and life wide perspective.

The paper is organized as follows. Section 2 presents several existing approaches and projects for the OLM and personal data. In order to identify key criteria in data management aspects. Section 3 details our prototype and demonstrates the feasibility of personal information management for Open Learner Models, based on an e-Portfolio example. Section 4 concludes the paper and presents its perspectives in the lifelong learning field.

2. RELATED WORKS

In this section, we consider existing projects, concepts, and approaches related to the OLM context and the personal data. That is why we present the TenCompetence project, the Army Learning Concept, the MyData-Midata projects, and the Learner Models/ Badges approach.

The European Network for Lifelong Competence (TENCompetence) Development is a European project aiming at developing an integrated open source infrastructure that enables and fosters lifelong learning \[12\]. Users are able to integrate, manage and carry out their competence development activities and their own competences in interaction with other users, through a Personal Competence Manager \[13\]. However, this manager is not directly connected to any source, and as the system was not fully deployed nor term access neither data disclosure are provided.

The Army Learning Concept (ALC) 2015 describes a learning model that leverages peer-based learning \[3\]. According to ALC 2015, the e-Portfolio is the central Learner Model that collects data from multiple sources and it is considered as an Independent Lifelong Learner Model. In this approach, learners owns data, however, the institution is considered as the steward of their OLMs, limiting the focus of the e-Portfolio to institutional aims, and not allowing users to claim their data.

The Mydata project in US (and another similar project named Midata in UK) works with businesses to give learners better access to their electronic personal learning data that companies hold about them. This is proposed in a broad context of data disclosure and openData. Those initiatives are steps in the right direction but neither giving access to the whole sets of data collected during learning nor providing specific services for managing those data. Note that these two projects are limited to specific countries.

The Open Badges is a concept proposed by the Mozilla foundation and is presented in \[14\] about an evidence-based source for Learner Models. The foundation proposes an infrastructure to create badges, to validate them through external (institutional) servers, to collect multiple user badges in a single backpack. Badges acquired may be published in social networking services like LinkedIn. However this backpack is not sufficient to manage badges according to personal goals.

These projects, concepts, and approaches enable us to define important criteria in the OLM context that allow the user control/persistence, and enable privacy as well as self-defined sharing:

- Data access: The solution must allow access to learners' data. The learner must be able to interact with data, to classify his Learning achievements and related outcomes, or to update some information.
- Data duration: Learners must be able to control data duration (long and short) according to lifelong goals. Nowadays, duration is defined by institutions according to their own requirements and ethics policies. Learning traces can only be stored for one year and used for predefined purposes. Student grades are generally stored during 5 years in French institutions. Learner himself may be interested in comparing current practices with his learning outcomes so long ago, or visualize long term trend indicators.
- Personal data storage: learners must have their personal data storage as a personal resource for organizing, preserving, sharing, and orchestrating personal learning data.
- Data transfer: learners must be able to download data from any platform, to upload and reuse them in their personal data storage. Two modes of transfer should be achieved: results (such as grades or diplomas must be provided by educational institutions) and learning traces collected along the learning process must be captured.

In the rest of the paper, the term “data management” refers to the data access, data duration, personal data storage, and data transfer.

Across the table 1, we found that no projects or approaches meet these criteria. That is why we have not been able to retain any existing approach and we need to propose a new solution covering all the criteria described above.

<table>
<thead>
<tr>
<th>Data access</th>
<th>Data duration</th>
<th>Personal Data Storage</th>
<th>Data Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>TenCompetence</td>
<td>Yes</td>
<td>Short</td>
<td>No</td>
</tr>
<tr>
<td>Army Learning Concept</td>
<td>Yes</td>
<td>Short</td>
<td>No</td>
</tr>
<tr>
<td>MyData – Midata</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Open Badge</td>
<td>No</td>
<td>Long</td>
<td>No</td>
</tr>
</tbody>
</table>

To highlight all these ideas, we are going to detail in the next section our approach that meets all these requirements and provides innovative solutions in this domain.
3. A PROTOTYPE OF PERSONAL INFORMATION MANAGER FOR E-PORTFOLIO

In this section, we present a proof-of-concept prototype based on a Personal cloud infrastructure and standard interface implementation to collect data. We demonstrate how our solution provides the required data management.

We use the Cozy cloud framework\(^1\) to implement the concept of Personal cloud. It provides a data oriented platform, with privacy and user control as key concepts. We choose this framework because it includes required components/functions: controller to manage applications, proxy to authenticate requests from users, and redirects them; and Data System, to store data and make sure applications only access the data they are allowed to.

Figure 1 shows the prototype architecture\(^2\). Infrastructure components of the personal cloud are highlighted in green. Learning components (including services) are highlighted in purple. In our architecture, e-Portfolio (1) is seen as an example of the Learner Model. It enables data access according to lifelong personal goals. Implementing such Learner Model in a Personal Cloud provides personal data storage (2), enabling full data access to the learner and full duration control as well. Data are collected in two different ways: external learning achievements may be collected through a data transfer mechanism (3) from external servers, whether institutional or commercial, or learning traces through a learning streaming flow (4). The proxy mechanism (5) provides a basic mechanism to grant access selectively.

In this context, we developed two data transfer connectors. The first data transfer connector retrieve Open-Badges, where the user may synchronize her personal learning achievement database with existing backpack. As validation of badges is maintained by external (institutional) servers, the user is only able to classify which ones are relevant for what purpose in her e-Portfolio. Other digital diplomas can be retrieved in a similar way. The second data transfer connector retrieve commercial e-Portfolios, the commercial e-Portfolio service provides a specific API enabling download of existing learner certifications. This service can be extended in the case of LinkedIn.

Once the data transfer connectors are implemented, we need to aggregate data from various learning sources, this must be achieved through specific API, based on linked data to enable higher semantic information level, or data streams. Those data are collected in data stores, providing access to various services see [15] like reflection, visualization, adaptive learning…

---

1. [www.cozy.io](https://www.cozy.io)
Learning Record Stores (LRS) to provide data access. Those standards are widely adopted in the open learning environments [16]. In our context, statements are duplicated in the learner personal cloud and the external LRS, enabling data collection for personal (4.1) and institutional (4.2) record storage at the same time. This gives the opportunity to fulfill institutional analytics needs, and also give direct access to the user. Our architecture also enables the exchange between personal and institutional records.

We developed a specific Learning Record Store compatible with cozy framework and based on the xAPI to enable data aggregation from various contexts. As it is embedded in cozy context, it ensures the user control, as well as the ability to fine grained control access to third party services and to other LRSs as well.

This prototype is able to store statements from various applications proposing a xAPI wrapper. We used some basic examples, and developed a specific wrapper we tested on nQuire, which is a personal inquiry learning system proposed by the Open University [17]. As a proof of concept, this wrapper sends activity statements to the user’s personal LRS and in parallel to his institutional LRS.

Considering our introductory example, thanks to our architecture, Alice can aggregate learning data from multiple sources, she can have in her personal cloud her diplomas from the primary / high school in Australia, from the school engineering in Belgium, her certificate of training from Germany, and other online certifications (Open Badges, LinkedIn, DoYouBuzz...). She can store for her whole life these certifications independently from any institutions or commercial platforms. She can also monitor ongoing activities through xAPI collection, including informal learning. She can manipulate her portfolio as a completely Independent Open Learner Model.

4. DISCUSSION AND PERSPECTIVES

This study addresses the problem of the learner data management in lifelong learning. The main questions of the study are how to render learners to be masters of their learning achievements independently from any platforms and what is the suitable functional and technical infrastructure to achieve this goal. We investigate the problem from its theoretical background, and we consider existing approaches for the OLM context and the personal data in order to see if any existing approach can meet our requirements. As shown in the state of art, no one can fully respond to our needs in terms of the support of the learner control, the privacy and the lifelong storage of data.

To achieve this, we propose an architecture that is a Personal Information Manager System for e-Portfolio that provides lifelong data access and storage aggregated from different sources, including learning achievements and learning traces. This solution provides a self-learning trace storage that can be deployed in collaboration or in parallel with other tracing systems.

Our perspective is to add the “exchange” dimension to the proposed architecture. This dimension, as described in the Alice’s scenario, is about exchanging her information with alumni association, initiating new collaboration based on her learning achievements and having feedback about her OLMs through interaction with her social network.

We are interested in the trust and scrutability dimensions. Trust is about supervising who use data and how they are used. Scrutability [18] is about understanding how the system arrived at the information the user sees. Both dimensions are ensured at the learner’s community level through the open source implementation of our solution.

Another important point is the evaluation of learner’s acceptability of our approach from the user-centric vision and the personal learning data vision. In our institution, learners are reluctant to use existing e-Portfolio platforms because initial investment is high and long term accessibility is unknown.

The question is to know if such independent e-Portfolio, with improved data collection and long term data duration may be more accepted by users and relevant for lifelong learner reflection than an e-Portfolio proposed by educational institutions.

We think that our proposed solution provides an open environment for innovation around lifelong and life wide learning. New services must be developed in order to test how the lifelong learners’ data management can promote the learning-to-learn skills (learning reflectivity, facilitate planning, enhance user participation, and monitor learning). This evaluation must be conducted by the learner’s community, the central actor and the final user, who must express their own needs in order to foster the feeling of community among all learners. Teachers and Institutions could be of course part of the evaluation, but would no longer be the leaders of learning services.

5. ACKNOWLEDGMENT

This work and submission are funded by the European ITEA3 13043 MOOCTab project (http://mooc tab.com/).

6. REFERENCES


Tracking and Reacting to the Evolving Knowledge Needs of Lifelong Professional Learners

Oluwabukola Mayowa Ishola, Gord McCalla
Department of Computer Science
University of Saskatchewan, Saskatoon, Canada
bukola.ishola@usask.ca, mccalla@cs.usask.ca

ABSTRACT
Our research is part of an ongoing project to provide tools to support the individual learning needs of lifelong professional learners. In the advanced learning technology research community there is an increasing interest in personalizing learning technology according to the evolving knowledge needs of the learner and the changing knowledge within their profession. In this paper we propose an approach to supporting the lifelong professional learner that adapts as the learner and the knowledge base itself change. The novelty of our approach is threefold. First, we use data from social media to gain insight about a professional learner’s knowledge, in particular to diagnose the gaps in their knowledge. Second, we don’t just diagnose what learners know and don’t know, but we also try to determine what they know about what they know and don’t know. Third, we track how the domain of expertise is itself changing. Ultimately our goal is to build an open learner modeling system wherein the gaps in the knowledge of professionals can be indicated to them at any point in time while providing personalized help also. In this paper we describe the architectural design of this system.

Keywords
Lifelong learning; professional development; diagnosis; learner modelling; knowledge states

1. INTRODUCTION
In the advanced learning technology research community, there is increasing interest in lifelong learning and in particular in tracking the evolution of both the learner and the knowledge to be learned [19]. Rapid technological advances are leading to massive ongoing change in society and work, driving the need for lifelong learning of the new skills and knowledge needed to succeed in this changing world [20]. As knowledge evolves, learners will need to continually update their knowledge and skill to effectively participate in continuous vocational and professional development [22]. Professional learning is an important subset of lifelong learning and a burgeoning area of advanced learning technology research [1]. Traditionally, the majority of support provided to professionals by their organizations is oriented around their specific job role, which might not necessarily keep the professional’s knowledge up to date with broader developments in their profession. We would like to support such a lifelong professional learner.

In our research we drew on ideas from the ecological approach to learning systems [21], wherein vast amounts of data about learners and their interactions with the world are mined “just in time” for patterns that can inform pedagogical decision making.

Such data-driven, just in time modelling allows the learning support system to actively respond to changes in both the learners and the knowledge to be learned. In our work on professional learning our goal has been to mine the peer- peer interactions of software developers who are using the Stack Overflow (SO) online forum so that we can find gaps in the knowledge of these software developers. The Stack Overflow (SO) forum is a “question and answer site for professional and enthusiast programmers” [http://stackoverflow.com/1]. This online forum contains the questions and answers, profiles, badges, reputation scores, and other data of over 5.5 million users. There are over 30 million questions and answers. This is truly a large scale repository of information about programmers and their help needs.

Having found gaps in the knowledge of professionals, we envision building an open learning support system wherein these gaps can be recommended to them at any point in time while providing personalized help also. This would allow for learner reflection, planning and self- monitoring which could promote learners to take greater control and responsibility over their learning.

2. STACK OVERFLOW OVERVIEW
Figure 1 below illustrates a typical question and answer in Stack Overflow (SO). Users in SO can vote up or down questions and answers, as indicated by the up and down arrows shown in Figure 1. The person who asks the question can mark one of the answers given as accepted; this is signified by the check mark sign in Figure 1. All questions are tagged in SO to indicate the subject area the question falls under. A question can have a maximum of five tags since a question could be related to more than one subject area. For instance, the question depicted in Figure 1 is related to “ios”, “osx” and “swift”. The total up-votes and down-votes obtained by each user in all their posts is shown by their reputation score in their user profile. An overall view of each user is kept in their user profile that includes the popularity of the individual in the forum.

User Table.
The user table contains personal information about the activities of over 5.5 million users on SO. Evidence of the know-how of users is shown by their reputation score. The posted questions and answers are voted up and down by other community members depending on their usefulness. Some usage statistics of the user table are shown in Table 1. As can be seen about 88% (4,847,640) of the users have reputation values less than or equal to 50 while 0.000145% (8) of users have reputation value greater than 50000. The Stack Overflow reputation data fits a power law in which the

---

1 Stack Overflow is a publicly available dataset, and as such does not require ethics review for such data as there is no expectation of privacy.
majority of users have a low reputation score; the higher the score the fewer the number of users.

Information such as knowledge interactions between professionals, questions and answers, up votes and down votes received, badges earned, and tags used serve as input into the system. Our first results have been promising, and have informed the design of the architecture.

The five major components of the system consist of the Learner Model, the Evolving Knowledge Ontology, Knowledge Diagnosis, Social Filtering of the Diagnosis and the Open Learner Model. A diagram outlining the architecture is shown in Figure 2 below.

Figure 1. Illustration of a Question and Answer in SO (Adapted from http://stackoverflow.com/tour)

Table 1. User Table: Descriptive Statistics

<table>
<thead>
<tr>
<th>Description</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reputation</td>
<td>1</td>
<td>122</td>
<td>882096</td>
<td>2170.77054</td>
</tr>
<tr>
<td>Up votes</td>
<td>0</td>
<td>13</td>
<td>15011</td>
<td>156.66970</td>
</tr>
<tr>
<td>Down votes</td>
<td>0</td>
<td>1</td>
<td>66734</td>
<td>292.39685</td>
</tr>
<tr>
<td>Profile views</td>
<td>0</td>
<td>15</td>
<td>116945</td>
<td>619.30762</td>
</tr>
</tbody>
</table>

Tag Table.
Tags are used in SO when creating questions to depict the knowledge area of the question. The tag table contains all tags in SO used to label questions and answers. Out of over 44,000 tags in this table, only 445 tags have over 10,000 related posts while over 25,000 tags have fewer than 50 related posts.

Post Table.
The post table in SO is the biggest table and is where all questions and answers are stored. Currently there are over 30 million posts on SO. The content of this table includes post id, creation date, and number of post views, message body, owner user id, owner display name, last activity date, last edit date, tags, title, answer count and number of comment to answers.

3. SYSTEM ARCHITECTURE

This section describes the architecture of our system to support lifelong professional learners. While the overall system has not been implemented and tested as yet, we have run 3 experiments aimed at exploring how to inform the underlying learner model with information that has been mined from the Stack Overflow forum. In these experiments we explored how to diagnose a learner’s state of knowledge, how to measure the influence of peers, and how to predict future states of knowledge from past states. The source of our data is, as discussed, Stack Overflow.

2 These experiments have been written up in papers that are currently under review. Our purpose in this paper is not to describe specific experiments but to provide an overview of our goals and approaches for a system faced with the challenge of supporting lifelong professional learning. We feel this is an appropriate goal for a workshop paper, and we hope we have succeeded.

The detailed description of the functionality of each component of the system is discussed in the following sub-sections.

3.1 Learner Model

The learner model contains information about the knowledge interactions for each individual user, the questions asked, the answers given, up votes and down votes received, badges earned, reputation score, and last login information for each individual user (who we also will refer to as the “learner”). Posts (questions and answers) made by the learner would be used in inferring their knowledge interests, and this creates the possibility of tracking changes in these interests over time. These interests can be inferred from the tags used in posts by each learner. As the tags change, the inferred knowledge interests of the learner change dynamically along with them. All of these interests over time will be represented in the learner model, which therefore not only captures current interests, but also a record of how these interests evolved. The knowledge interests of the learner during a period of time period \( t \) would be determined by mining all tags employed in questions asked by the user within \( t \). We compute the tag distribution employed in question posts for each user as shown below.

\[
D(u,t) = (N_1/N_{total}, N_2/N_{total}, \ldots, N_i/N_{total}), \quad \text{where } N_{total} = \sum_i N_i
\]

The count of questions asked by learner \( u \) for tag \( i \) is represented by \( N_i \), while \( N_{total} \) shows the total number of questions asked by the learner for time period \( t \) for all the tags represented in learner model. Computing \( D(u,t) \) shows the tag distribution for each user for the time period \( t \). As the knowledge needs of the learner evolve from time period \( t \) to the next time period, we can compare how the...
knowledge interests of the learner differ within this period of time. In establishing this comparison, we compute $d_i$ similarly to Liu et al. [2]:

$$d_i(X, Y) = \sum_i |x_i - y_i| \quad \text{and} \quad d_\infty(X, Y) = \max_i (|x_i - y_i|)$$

$X$ and $Y$ represent the tag knowledge interest distribution of the learner for two time periods. $d_i$ represents the variation in the knowledge interest of the learner between the two periods, while $d_\infty$ represents the maximum variation in the interests of learners between the time frames. The smaller the value of $d_i$, the closer the similarity in the knowledge interests of the learner between the two periods of time. Tracking how the knowledge interests of a learner evolve over time would help in adapting support so it can be focused on the current knowledge interests of the learner.

In addition to tracking tags, the learner model also keeps a record of the learner’s questions and answers, which ones have been up voted or down voted, what badges have been earned, the learner’s reputation, and so on. These can be useful in interpreting levels of activity in posts, while Tag $A$ Tag $B$ represents the number of times Tag $A$ and Tag $B$ have been used in posts, and more specifically, Tags $A$ and $B$ with each of $git$, $commit$, $reset$, and $revert$.

3.2 Evolving Knowledge Ontology
This “evolving knowledge ontology” module of the proposed system would contain knowledge areas drawn from tags used in questions. As mentioned earlier, tags in SO help in classifying questions into different knowledge areas and also in identifying the questions whose answers would interest a user. While asking a question, it is possible to use at most 5 tags in SO. An example of a typical question in SO with tags assigned to it is shown in Figure 3.

![Figure 3: Sample Question in Stack Overflow](image)

As shown in Figure 3, the four tags used in creating the question are $git$, $git-commit$, $git-reset$, and $git-revert$. While four tags were assigned to this question, it could be inferred that this question is broadly related to $git$ and more specifically $git-revert$. This implies, looking at the tags used in posts, relationships between these tags could be inferred. For instance, $git-commit$, $git-reset$, and $git-revert$ could all be said to be related to $git$. This leads to the requirements of this module: not only to serve as a repository for all tags used in posts, but also, to build a tag ontology which could represent how all tags used in SO are related. Currently there are 44,917 tags existing in SO with the possibility of more being added as new knowledge areas emerge. These tags vary in their popularity among users in the forum. We aim to build a hierarchical tag ontology representation that would depict the relationship between tags. The parent-child relationship would be determined based on the co-occurrence of tags as used when creating posts. Jaccard Coefficient [28] would be used to calculate the co-occurrence of tags as shown in the formula below:

$$\frac{|Tag_A \cap Tag_B|}{|Tag_A \cup Tag_B|}$$

Tags $A$ and $Tags_B$ are the number of times $Tag_A$ and $Tag_B$ have been used in posts, while $Tag_A \cap Tag_B$ represents the number of times the two tags occur together in posts., Using the Jaccard coefficient (JC) computation above, the closer the JC between two tags to 1, the closer to each other they would be in the hierarchy. For instance, as shown in figure 3, the JC between “git” with each of “git-commit”, “git-reset”, “git-revert” and any other related tags that have been used with “git”, would be computed. The positioning of these tags for the branch of “git” on the hierarchical tag ontology would be determined base on the closeness of the JC to 1. Again, the envisaged tag ontology module would evolve dynamically over time. As new tags are being created, this information would be added to the tag ontology repository, while older tags with less popularity could be labelled as of fading interest. The rise and fall in the popularity of tags as represented in the tag ontology would also serve as a guide to identifying knowledge trends within the software development profession. We believe a forum like SO with currently over 5.5 million professional users and over 30 million posts, is an appropriate forum to infer knowledge trends within the profession. For instance, a new tag, which has attracted a huge number of views from users within period of time $t$, could represent a trending new topic within the community while an older tag with lesser views and usage from users within period $t$ could represent a fading topic.

3.3 Knowledge Diagnosis
Professionals, however well trained and experienced, often have gaps in their knowledge, and are often unaware of these gaps. Previous studies [3, 4] have classified the knowledge of a person into four possible “knowledge states”: the things we know we know, the “known knowns” (KK); the things we know we don’t know, the “known unknowns” (KU); the things we are not aware we know but we do know, the “unknown knowns” (UK); and, lastly, the things we don’t know we don’t know, the “unknown unknowns” (UU). Detecting the knowledge states is the goal of the “knowledge diagnosis” module, with a particular focus on determining the “knowledge needs”: the KU and UU that constitute “gaps” in the professional’s knowledge [29].

In diagnosing the knowledge states of each individual learner for each knowledge area represented in the ontology, we would employ the number of up votes and down votes received by the learner for each respective knowledge area. The “known knowns” would be determined by looking at the distinct answers the user has given under each tag that were up-voted. The “known unknowns” would be determined by looking at the tags of questions the user has asked. The “unknown unknowns” would be determined by looking at the tags of questions that the user has answered where the answer was down voted. As to the “unknown knowns” these seem less informative, so at this stage we have not sought heuristics to find “unknown knowns”.

In diagnosing the knowledge state of the learner for a given topic in the knowledge ontology at a given time period $t$, we simply count the number of KK, KU, and UU posts for “leaf node” tags in the ontology for a given learner and determine the relative percentage of each. The highest percentage exhibited by the user is diagnosed to be their knowledge state for the topic represented. For instance, a user whose evidence of KK for java is 70%, KU for java is 20% and UU for java is 10%, would be determined to know java, i.e. java is a “known known”. Once leaf node knowledge states are determined, these can be propagated to higher level nodes on a “highest percentage of children” nodes basis. Again, these knowledge states are added to the learner model, indexed by time period $t$. 

[28] Jaccard Coefficient
[29] Knowledge Diagnosis
3.4 Social Filtering of the Diagnosis

Having inferred the current knowledge interests and knowledge states of a learner, we want to determine what this implies about the learner’s evolving knowledge needs and possible ways the learner might be helped in meeting these needs. We do this using a social filtering approach. We compare learner u at time t to other learners with similar learner models at a given time period t₁ in the past. These similar learners are useful in at least a couple of ways. First, these similar learners could be a source of advice or help in overcoming learner u’s knowledge gaps, assuming that the similar learners themselves have done so (i.e. knowledge inferred as UU in the past is now in state KK). But more interestingly for our system, it is possible to look at what happened to the similar learners in order to predict what will happen to learner u going forward. Particular knowledge gaps may be seen to have been very important in generating lots of questions and confusion on behalf of the similar learners, while others may turn out not to have had much impact at all (with no further questions related to these knowledge areas). These insights can be used in helping to categorize and prioritize the knowledge needs for learner u.

This kind of collaborative filtering, common in recommender systems for example, allows our ecologically adapting open learning system to continuously be comparing any given learner to other similar learners, and “push” that learner forward in directions that have proven useful to his or her peers. All of this, in turn, is in line with our goals that the learning system evolves naturally along with the professional learners and their discipline itself.

3.5 Open Learner Model

The overall goal of this research is to inform professionals about their individual knowledge needs. The aim of the “open learner model” module of the proposed system architecture is to provide a support system that gives feedback to the professional learner about their knowledge states, an open learner model for them to peruse. This module displays to the learner their diagnosed knowledge needs, as defined by the detected gaps and socially filtered by comparison to learners who had similar knowledge gaps. The social filtering can allow inferences about what is important and not important. It can also allow inferences to be made as to an appropriate order in which the knowledge needs could be met (essentially what would be called an “instructional plan” in AIED), again based on what worked well for similar learners and in what sequence. There are a host of issues around how to do these inferences, how to display these to the learner, how to explain the nature of the gap to the learner, what kinds of interactions and control the learner will be allowed, and so on. These are the subject of current research. We are confident, however, that the Stack Overflow database is a rich source of insight about even these “HCI” issues, and perhaps even can directly supply content (for example recommending SO posts that explain the nature of a particular knowledge gap).

In the future we hope to be able to use similar techniques to find patterns in how other users have behaved in SO that would help to predict forgetfulness in the knowledge of the learner and thus allow us to be able to prompt the learner when evidence of forgetting arises. Also, in future we envisage the possibility of augmenting the online forum with other information about the professional: their resume or e-portfolio, their LinkedIn profile, the artifacts they produce (e.g. code), the tasks they have been assigned, job performance evaluations by themselves, their peers and their man-agers, etc. Multiple sources of knowledge like this would be a rich mine for further understanding of the knowledge states of individual users. Such sources would also offer the possibility of more refined personalized diagnosis, not only of KK, KU, and UU, but also of UK (where, for example, behavior in an online forum that indicates knowledge of various topics could be compared to an e-portfolio for topics not mentioned as known by the professional).

4. RELATED RESEARCH

Today, learning has become part of our daily life. Lifelong learning is a necessity for all of us, but particularly so for professionals, whose knowledge and skills are challenged by changes stimulated through the emergence of new technology [5]. The ongoing need to acquire knowledge transcends the walls of the classroom with a need to continuously improve skills, competence and knowledge [6]. The internet and the World Wide Web have contributed greatly to this need for lifelong learning, but provide new opportunities to support this learning as well.

Lifelong Professional Learning

From an advanced learning technology perspective, lifelong learning has created interesting areas of research touching upon personalization, collaboration, ubiquitous learning and much else [7]. Systems to support lifelong learning need to be able to adapt to the specific individual learning needs, preferences, gaps, and goals of each professional. Some important aspects of such personalization include the ability to reuse existing information across applications, to be able to do life-logging of the activities of professionals, and to support professionals’ self-monitoring and reflection [8, 9]. Issues such as forgetfulness, continual change in learning goals, and interoperability of models of learners across various devices and applications are challenges in effectively personalizing learning in lifelong contexts [8, 9, and 10].

Bartkowiak performed a user study to determine factors responsible for gaps in the knowledge of professionals [11]. Results obtained from this study show that employees identified ineffective communication, lack of practical experience, and poor business management as causes of competency gaps. Employers identified efficiency of staff, ability to combine theory and practice, and lack of experience in the organization as possible causes of competency gaps. This study concludes that high self-awareness and ability to apply theoretical knowledge to practical problems are important.

Ley et al. [12] measured gaps in professionals’ knowledge by comparing the previous tasks performed and tasks to be performed in the future. Ley and Kump [13] argued that the number of tasks performed is a weak measure in assessing the competency of professionals. Rather, qualitative differences in events are more effective in determining competency, and these can be captured in knowledge indication events (KIEs). KIEs also have limitations. First, emotional states of professionals could interfere with their performance [12], even down to their keystroke behavior [14]. Also, collecting too much detailed information about a user could pose a challenge, as the data would grow in geometric proportion, even though only a small fraction of the information collected would be useful for adapting and personalizing the system [15].

As in this other research, we are trying to detect gaps, but rather than looking at knowledge indication events based on tasks performed, we instead examine the online behavior of professionals as they interact with one another in an online forum (Stack Overflow), seeking evidence of what they know and don’t know. We are especially interested in their unknowns, the “gaps” in their knowledge: both their “known unknowns” (KU) and their
“unknown unknowns” (UU). As discussed we hope to be able to diagnose the gaps in professionals’ knowledge in order to build learner models that could inform the professionals themselves of their knowledge gaps and how to overcome these gaps.

Recommender Systems
Some of our techniques, as discussed above, are drawn from recommender systems. Recommender systems have gained increasing popularity over time both within the learning community and in industry [16]. The use of recommender systems in the education domain differs from other domains, meaning that some traditional techniques can be adapted while others cannot [19].

The application of recommendation techniques in learning systems has been geared towards adapting learning resources to learners based on their learning preferences and preferences of past users [23, 24]. Also, it has been applied to modelling individual differences between students, so that the learning software can be personalized according to each student’s interest [25]. Zaiane [26] applied web-mining techniques to build an agent that could recommend online learning activities or shortcuts to learners in web-based course. The recommender agent consisted of the “learning” module that learns from past learners’ activity and the “advising” module, which applies the learned module to offer recommendation to students. Heraud et al. in their work provided contextual help to learners by adapting their learning session in providing link structure for the course [23]. Social relationships among individuals have been studied by mining social networks in ITS using collaborative filtering where recommendations are made about the interests of a user based on the preferences of other users with similar tastes [24]. This is similar to our use of collaborative filtering, although we are interested in finding gaps and we are reasoning over a much more complex learner model. Further, as discussed in section 3.4, we hope to be able infer pedagogically useful sequencing information (simulation experiments in our lab have already shown that this is potentially possible) [27].

Tang and McCalla in 2003 proposed an evolving web-based system that can adapt itself to learners and to the open web. Building on this study, Tang et al. employed learners’ interests and accumulated ratings given by other learners in recommending learning resources [17]. The success of this study was its ability to go beyond the confinement of closed learning environments by extending the recommendation of learning resources to include the open-web. Even though this work had similar goals to ours, it was fairly small scale, aimed at learning a known curriculum, and not designed for lifelong learning. The research never contemplated an actual system using a noisy, real world environment such as Stack Overflow that would be active for years.

Although, recommender systems have already gained prominence, the “cold start” problem of building the initial data needed for recommendation remains evident [16, 18]. MovieLens.org, a recommender system for movies addressed this problem by asking new users to rate their preferences for movies before the system can provide recommendation. This solution does not apply to learners, who in most cases are not able to rate learning artefacts in advance as they might not have sufficient prior knowledge. However, cold start isn’t as big a problem in our context as in other contexts, since there are millions of users with which to compare any given user. New users can be mapped to other new users from the past, and after only a few interactions with SO important insight can be gained about how the new user compares to the many users who have gone before him or her.

5. DISCUSSION AND FUTURE WORK
The competency of professionals has been determined in the past mainly by tracking their job performance [18]. This is not sufficient to judge their overall competence in their profession since the specific job (and workplace) will likely require only a subset of the skills they need to be fully capable professionals. Our approach to diagnosing their knowledge needs by comparing their competency with their peers, would allow professionals to see if their skill is rising or falling in comparison with others in their profession. Diagnosing their specific knowledge states would allow the lifelong professional learner to identify specific strengths as well as to identify gaps in their knowledge of which they might not even be aware. Even as the knowledge within the profession evolves over time, so also do the learning interests of the learner. The recommendations proffered to the learner would likewise evolve. Adapting learning to continuous change in knowledge within the profession is vital in keeping the learner up-to-date with the current knowledge states.

Even though (as mentioned in the introduction) we have carried out 3 experiments in which we have explored various aspects of this approach (including mining the data of hundreds of thousands of SO users), we, of course, need to do further implementation and evaluation, ultimately of the full architecture. However, in a workshop context we wanted to present a strong argument at the conceptual level for our approach in order to stimulate discussion. We feel that this research, even in its current “in progress” status, is interesting and original in its arguments, especially for the use of social media as a major source of insight about professional learners’ knowledge; in the use of knowledge states (KK, KU, and UU) that emphasize the awareness of the professional about their knowledge not just the knowledge itself; and in being designed for a noisy real life lifelong learning context. Perhaps most importantly our approach is ecological and evolutionary in the sense that it naturally evolves as the world changes, and is potentially capable of tracking changes in professional learners as well as changes in the profession itself without the need for a massive ongoing knowledge and software engineering effort. We look forward to a vigorous (and hopefully constructive!) discussion.

6. ACKNOWLEDGMENTS
We would like to thank the Natural Sciences and Engineering Research Council of Canada and the University of Saskatchawan for providing funding to support this research.

7. REFERENCES


