Exploiting Readily Available Web Data for Reflective Student Models

Judy KAY and Andrew LUM

School of Information Technologies,
University of Sydney, NSW, 2006, Australia

Abstract. This paper describes our work towards building detailed scrutatable student models to support learner reflection, by exploiting diverse sources of evidence from students use of web learning resources and providing teachers and learners with control over the management of the process. We build upon our automatically generated light-weight ontologies. In this paper, we describe how these support inferences from the fine-grained evidence that is readily available to higher level learning goals. To do this, we have to determine how to interpret web log data for audio plus text learning materials as well as other sources, how to combine such evidence in ways that are controllable and understandable for teachers and learners as required for scrutability and finally how to propagate across granularity levels, again within the philosophy of scrutability. This paper reports our approach to these problems. We report evaluation of this approach. This is based on a qualitative usability study, which indicates that users demonstrated good, intuitive understanding of the student model visualisation with system inferences.

1. Introduction

Student models have one obvious role as the drivers for personalisation. Importantly, externalised or open student models have another invaluable potential role, in supporting learning by helping learning by enabling improved learner reflection [1]. They also can be a useful basis for feedback to the teachers [2]. We would like to teachers to enhance their web-based or web-enhanced courses with learner models that are useful for reflection. This means that the processes of building the learner models need to be tailored to typical classroom teachers, being understandable and quick to use. To make the models useful for reflection, they must model the learners at varying levels of granularity, coarse grained so that learners can see how they are doing on the overall learning goals and fine grained so that they can determine which elements of work contribute to this higher level goals. Moreover, we want learners, and teachers, to feel in control of the modelling and to be able to scrutinise the models, delving into details of the processes that determine the model.

Web-based and other interactive learning systems differ from typical classroom learning in that they can easily provide very large amounts of data about learner. Unfortunately, that data is typically of very poor quality, as for example, in the case of detailed logs of page visits, time spent on each page and links selected. This is low quality in that a visit to a web page is weak evidence that the user read the material, let alone learnt it. On the other hand, if learners have not even visited the web pages for a course, it seems likely that they are less likely to have learnt the material than would generally be the case for learners who have
visited the page and displayed it long enough to have read it. Evidence of this sort is so readily available that it would be valuable to find ways to exploit it to build student models. Web learning environment also may provide higher quality evidence about learners. For example, there may be marks for class exercises, results of on-line quizzes and multiple choice questions. Evidence from all such sources tends to be fine grained in the sense that a single page of an on-line course is about a small part of it and a quiz question or set is typically about a current, specific sub-topic.

We want to support scrutable learner modelling which exploits the combination of the full range of types of evidence available. This poses several challenges. First, we need to determine how to interpret the evidence available. For example, we have a course with on-line lectures, each composed of a series of text slides with audio content. We need to determine how to interpret the evidence that a student attended such a lecture. Secondly, once the evidence is available, we need to determine how to combine diverse evidence sources, a task that has been the subject of a substantial body of research, including for example [at least 3-4 references, ideally overviews, mislevy, Dempster shafer, bayes], but we want this process to be readily controllable by learners and teachers and to be scrutable so that our system can provide simple explanations about how it the modelling works. A third problem is to be able to reason from the fine-grain level of the available evidence to the coarser grained higher level concepts. Our approach to this is to exploit an existing tool, Mecureo [ref], which builds lightweight ontologies automatically by analysing subject-area glossaries. This approach is very attractive in relation to our goal of scrutability because the dictionary is then a useful resource for providing explanations of the ontology: we can simply explain that the reason the system treats two concepts as related is based upon the relevant dictionary definitions, which can be presented to the learner. This approach also meets our goals of low cost construction of student models since it defines a structure for the user model automatically. There is an established body of work on ontological inference [refs] but we need a mechanism that is easily explained and intuitive and which can operate on our light-weight Mecureo-generated ontologies.

This paper describes our work towards tackling these challenges. Section 2 outlines our approach and Section 3 discusses the evaluation framework and infrastructure. Section 4 presents the results of a user study and Section 5 concludes with related work and discussions.

2. Dealing with diverse evidence sources and fine grained evidence

To tackle the problems of varying quality of evidence from different sources and varying amounts of evidence, we introduce the notion of a Student Standard. In the case of a course or teaching system, the Student Standard may be defined as the teacher sees fit: for example, a teacher in a mastery-based course may define it as the student model of the student who earns full marks for assessments and a perfect attendance record by the end of the course. By representing user model values relative to that of the Student Standard, students immediately have a gauge of their own performance and a visible goal to work towards. This is similar to the approach in [4] where student progress is mapped as an overlay to a directed graph that represents an idealistic student’s learning path. In our case, we model against the student who performs all available tasks exactly as the teacher intended. Note that this is not the same thing as an expert. Moreover, the teacher may well define the Standard Student as they choose. The model associated with the Student Standard should be easy for an instructor easy to compute in a closed course with fixed sources of evidence for learning, such as is commonly true for courses with on-line resources and activities as well as off-line class assessments.
We use ontologies to structure the user models, giving us the power to link the concepts in a meaningful way. The ability to reason about higher level concepts and also find related concepts is driven by the ontology. At the same time, the ontological structure can be exploited by visualisations to provide a suitable interface to the user model [3]. However, formal ontologies are difficult and time consuming to build. It would be hard to imagine an average teacher undertaking such an enterprise (and then they’d have to mark up all their learning material accordingly).

On the other hand, light-weight ontologies are often cheap to build using automated means, but lack the structure that would make it suitable to use formal reasoning techniques. Instead we must use other methods.

2.1 Reasoning about Fine Grain Knowledge

Fine grained evidence comes from a variety of sources, but most commonly from assessments (such as tutorials, assignments, projects and exams) as well as access to resources (for example lectures and background readings). Fig. 1. shows an example user model where evidence feeds mainly into the fine-grain concepts.

![Fig. 1. A student model with fine-grain evidence for learner knowledge of concepts in the HCI domain. Evidence may feed into a single concept (such as sources 1, 3, 4 and 5) or multiple (source 2). Evidence may also feed into higher levels of the ontology, in this case, source 4 feeds into a mid-level concept rather than a leaf concept. They may also come from different sources. In this case, some evidence is from web log data (sources 1, 4 and 5), and some is from hands-on laboratory marks (sources 2 and 3).

There is a challenge in the fact that the number of evidence sources varies for different concepts. Obviously, concepts with less evidence sources will on average have less evidence, which poses a problem for the reliability of the concept value. For example, concept A may have four web pages and a tutorial to teach it, yet concept B may only have two tutorials. This may be due to reasons related to the layout of the learning materials, the organization of the teaching or other pedagogical issues. One solution is to use a comparison to a static user model, i.e. the Standard Student model in the case of a learning course, for each concept during resolution, so we end up with a relative measure rather than an absolute one. Because every value becomes a relative value in relation to the Standard Student, we reduce the effect of the varying amounts of evidence for the concepts.

We now explain this approach, as we’ve refined it, in a system with two types of evidence: the amount of time students spent listening to audio for online learning objects, and the marks they received for weekly tutorial sessions.
For the audio evidence, the length of audio narrative for each slide is known. We assume the Standard Student will have listened to the full slide (and have an extra bit of leeway time for taking notes, etc). We can compare the length of time a user has spent on each slide to that of the Standard Student time, and assign a score based on this. The weightings we assign range from 0.0 to 1.0 and the breakdown are shown in Table 1.

<table>
<thead>
<tr>
<th>Understanding</th>
<th>Duration on slide as percentage of Standard Student Time</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seen</td>
<td>Student Time &lt; 10%</td>
<td>0.1</td>
</tr>
<tr>
<td>Partial Heard</td>
<td>10% &lt;= Student Time &lt; 80%</td>
<td>0.5</td>
</tr>
<tr>
<td>Standard Student</td>
<td>80% &lt;= Student Time &lt; 150%</td>
<td>1.0</td>
</tr>
<tr>
<td>Overheard</td>
<td>150% &lt;= Student Time</td>
<td>0.8</td>
</tr>
</tbody>
</table>

The Overheard weighting is slightly lower than the Full Heard. This is to account for the times when students have become distracted with other activities and have left the browser open. All of the values from each audio evidence source for a concept are then averaged. This results in a final value from 0 to 1.0; a perfect student will have listened to every slide as a Full Heard, resulting in a value of 1.0 for the component. We call this the Normalised Audio Score.

For the tutorial evidence, the students receive a mark out of 10. A perfect student should get full marks for every tutorial in our course, so in effect a mark out of 10 is already a comparison against that of the Standard Student. We sum all the tutorial evidence scores for a particular concept and divide by the total possible marks (Standard Student’s score) to get a value between 0.0 and 1.0 for the final value for tutorial evidence. We call this the Normalised Tutorial Score.

To combine the two values, we use a simple formula to determine each evidence type’s contribution to the final score:

\[
\text{Score} = k1 \times (\text{Normalised Audio Score}) + k2 \times (\text{Normalised Tutorial Score}) \text{ where } (k1 + k2) = 1. \tag{1}
\]

Based on an intuitive sense of the reliability, \(k1\) has been set to 0.25, and \(k2\) has been set to 0.75 when there is tutorial evidence.

2.2 Reasoning about Coarse Grain Knowledge

We need to be able to model about the user’s knowledge of higher level concepts. We want to deal with the case where there is no direct evidence at all. For example, in Fig. 1., there is no direct evidence for the concept \textit{usability} as evidence sources contributes to concepts finer grain.

Fortunately an ontology structure can provide a structured granularity to the domain concepts. Granularity is useful as we can focus on particular contexts in the domain to reason about [5]. For example, in Fig. 1. we can consider the subtrees spanning from \textit{Empirical} and \textit{Predictive} as two separate smaller contexts to reason about. We want to infer the level of student knowledge about \textit{predictive} usability, based upon the fine-grained evidence for \textit{heuristic evaluation} and \textit{cognitive walkthrough}. We also want to be able to infer the student's knowledge about \textit{empirical} evaluation from evidence about methods such as \textit{think aloud}. In addition, we want to be able to infer about the learner's knowledge in the broad area of \textit{usability}, based on their knowledge of both \textit{predictive} and \textit{empirical} usability.
evaluation. Given an ontology and a user model built from evidence for the fine grained elements, we need a mechanism to reason about the coarse bgrained core knowledge.

One simple method is to do a spanning tree from the leaf concepts (the fine grain) and recursively pass their values up to the parent concept till we reach the higher level coarse grain concept we want to reason about. At each stage when the values are passed up the tree, some calculations can be done to factor in the distance from the coarse grain concept in the tree, as well as the amount or type of evidence.

An example of this is the averaging model we present below. We can recursively run this algorithm up the tree till we reach the root concept we are inferring about.

For a particular concept, we take an average of the values of the child concept values. This value is then multiplied with \((1 - \text{value of root concept})\) and added to the value of the root概念 to give a proportional boost, but always maintaining a value between 0 and 1. The lower the score of the root concept, the higher proportion of inference the value will take. Equation (2) summarises the averaging formula for a concept \(v_a\) with \(n\) related concepts, where \(n \geq 1\). In the case of \(n = 0\), \(v_a = v_a\) (i.e. there is no inferred contribution to the final score for this concept).

\[
\text{Equation (2)}
\]

Consider the example portion of a student model shown in Fig. 2. We want to infer about concept Usability. We suspect that the student should understand more than what the direct evidence of 0.1 indicates, since the two related concepts (Predictive and Empirical) seem to be well understood by them. We substitute these values into formula (2) and arrive at the value 0.65 as the new value for Usability – a quite reasonable assumption based on the knowledge of the fine grain concepts (3 & 4).

\[
\begin{align*}
    v_{\text{usability}}' &= v_{\text{usability}} + (1 - v_{\text{usability}}) \times \frac{v_{\text{predictive}} + v_{\text{empirical}}}{2} \\
    0.65 &= (1 - 0.1) \times \frac{0.6 + 0.4}{2}
\end{align*}
\]

Fig. 2. A portion of a user model where the coarse grain concept Usability has a low score, despite the fact that the partitive components both have high scores. It would be useful to be able to infer more about the coarse grain concept from the fine grain concepts Predictive and Empirical.
3. Evaluation Framework

The User Interface Design and Programming course taught at this university is the demonstration environment for the tools and also the evaluation domain. It has 241 audio-slides (lectures are a collection of visual slides with audio narrative). There are also live lectures and laboratory classes. For the evaluation, we used the subset of material about design and HCI (161 slides organised in 9 lectures).

We now describe, very briefly, the process used to build the student models. This draws upon several tools that we have constructed:

- Mecureo [6] to construct the domain ontology;
- Metasaur [13] to link each learning object with metadata concepts from the ontology;
- Personis [14] to represent the student models;
- slide-evidence extractor which analyses web log data to create evidence based on student accesses to the slides;
- tutorial evidence extractor which uses tutorial marks to create evidence;
- SIV (Scrutable Inference Viewer) [8] to provide the interface for users in the study.

The domain ontology built by Mecureo [6] was automatically from the Usability First Glossary [http://www.usabilityfirst.com/]. This has 1,129 terms and categories. Mecureo analysed the dictionary definitions to construct a lightweight ontology based on the relationships between concepts defined in the dictionary. We augmented these with 105 additional definitions, giving a total of 1,234 concepts and 10,690 relationships between them.

We used the Metasaur interface [13] to create metadata, based on the set of concepts in the domain ontology, associating these with each lecture-slide and tutorial [7]. Importantly, this subset of the ontology's concepts automatically defined the components of the student model definition in the Personis user modelling representation [14]. This subset of the ontology has 190 concepts and the 423 relationships between them. The tools that collected evidence from web accesses and tutorial performance were used to add evidence to each student's learner model.

The reasoning methods described above operate as resolvers in Personis. The result of this process is available for the learner to scrutinise, with the Scrutable Inference Viewer [8] (SIV) interface. This provides an interface for visualizing the user model and to scrutinise the basis for what was displayed. Fig. 3 has a screenshot and explanation of its elements. For the evaluation, only the concepts that appear directly or are high level in the course are included in the visualisation.
Fig. 3. Example SIV interface. The visualisation is at the left, with the concepts listed vertically. The concept *user interface critique* is in focus and has the largest font. Related fonts are in the next largest fonts, and unrelated concepts are blurred out. Horizontal position indicates the amount of evidence for that concept in the user model. Concepts with a score greater than 0.5 are in green, others in red. The list of evidence contributing to the concept score is at the right – in this case there is no tutorial evidence, and the score for the concept is 0.86. The inferred evidence is determined using the averaging formula (2).

4. Usability Study

We used a think-aloud evaluation. Participants were six senior level undergraduate students, all with experience as teaching assistants. They were asked to take the role of tutors and were presented with the information sheet below. They were asked to think-aloud as they performed the task in Fig.4. In particular, we were interested in whether they:
1. could use the interface and understand it
2. would consider the results of the inference reasonable
3. could see the related concepts contribute to the reasoning

\[ \text{Concept: user interface critique (0.86)} \]

Audio Evidence (raw 0.80, contribution 0.80)

This has been inferred from the following evidence:

- The lecture data Prof145ta1t1 was attended for a duration of 112 seconds.
- The lecture data Prof14ta9 was attended for a duration of 211 seconds.
- The lecture data Prof145ta1t1 was attended for a duration of 599 seconds.
- The lecture data Prof145ta1t1 was attended for a duration of 312 seconds.
- The lecture data Prof14ta9 was attended for a duration of 211 seconds.
- The lecture data Prof145ta1t1 was attended for a duration of 599 seconds.

Tutorial Evidence (not present, contribution 0.0)

- There is no tutorial evidence for this concept.

Inferred Evidence (contribution 0.06)

- This extra contribution has been inferred from the terms visible at depth 2.

2 Colour screenshot at http://www.it.usyd.edu.au/~alum/assets/screenshots/siv-um05-1.jpg
Students A and B have quite different competence for the User Interface Design and Programming course. The course coordinator has requested that students struggling in this area will be invited to attend an additional catch-up tutorial session.

As a tutor for the course, you want to see how well the students understand the concepts in the area of predictive usability, in particular the concepts cognitive modeling, heuristics and user interface guidelines. You need to fill out a form to allow them to attend the tutorial session as there is a limited number of places.

Unfortunately there is little direct evidence for these concepts, though there are plenty of more specialized concepts (such as the fact they have listened to a lecture on cognitive walkthrough, which is a subtopic of cognitive modeling) with evidence that could contribute to their understanding of the concepts you are after.

You want to select these topics on the signup sheet (and maybe some additional ones) relating to this area of study and see what the system infers about the student’s knowledge.

Decide if Student A and/or Student B should attend the catch-up tutorial with a justification for why they should attend on the signup sheet.

Fig. 4. The task description for the evaluation given to the participants.

Two pseudo-students, A and B, were created, both based on a real student at the middle of the class ranking in the User Interface Design and Programming course. They were identical, except that student B had failed to attend several online lectures, and so had no web data for these. In addition, student B had lower tutorial marks than student A. Using SIV inference for course grain concepts, B’s scores were consistently lower than A’s.

The three concepts, cognitive modeling, heuristics, and user interface guidelines all had no evidence; hence a resolved score of zero in the user model, resulting in bright red font and, as these had no evidence, they were at the far right in the visualisation. Table 2 shows the values for the three concepts after inference based on evidence for related concepts. Student A’s higher degree scores for fine grain concepts is also reflected in the inferred values.

Table 2. The inferred values for each concept

<table>
<thead>
<tr>
<th>Concept</th>
<th>Student A</th>
<th>Student B</th>
</tr>
</thead>
<tbody>
<tr>
<td>cognitive modeling</td>
<td>0.50</td>
<td>0.22</td>
</tr>
<tr>
<td>heuristics</td>
<td>0.87</td>
<td>0.23</td>
</tr>
<tr>
<td>user interface guidelines</td>
<td>0.62</td>
<td>0.33</td>
</tr>
</tbody>
</table>

All the participants successfully completed the task and from the results in Table 3, unanimously decided that student B should attend the extra tutorial session.

All participants started with the search tool to look for the topics and quickly correlated the colour of the topics with the degree of knowledge for the students. All participants based their judgment student B’s poorer understanding compared to student A because student B’s inferred scores were all lower.

Table 3. The information written by participants on the signup sheet

<table>
<thead>
<tr>
<th>Participant(s)</th>
<th>Student</th>
<th>Reason for attending extra tutorial session</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>They do not have a good understanding of the above 3 concepts.</td>
</tr>
</tbody>
</table>
Table 1:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>B</td>
<td>Although there is no direct evidence of the student’s understanding of the three concepts, by inferring other concepts that are related to the three concepts, probability of the student understanding the concepts is low.</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>Inference readings returned low as no data on many of the related topics.</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>Although there is no direct evidence in the form of audio/video evidence of student A or B understanding the concept. The inferred evidence based on the relationships or underlying concepts suggest that student A has more knowledge than student B as the values for the inferred evidence are higher for all three concepts.</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>Need more details and info on these topics.</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>Low inferred score for all 3. The concepts looked red all the time.</td>
</tr>
</tbody>
</table>

Some pointed out upon seeing student B’s user model that they were not as good as student A based simply on the distribution of the colours when the concepts were expanded. Participant 5 said for the concept *user interface guidelines*, “In this case, there’s more greens for this topic for student A [than student B]”.

These comments were made before the participants used the Infer button to see the inferred value. They seemed to be happy that the inferred values matched their expectations. Participant 1 selected cognitive modeling for student A and instantly said “Cognitive modeling comes up red. I infer because the other concepts are green”. For student B on the same topic, Participant 1 stated “cognitive modelling appears correct [coloured red], but I will infer to make sure”.

Participants could also equate the inferred value with the values for related concepts. For example, Participant 6 was asked if they could see why the inferred value for heuristics indicated that Student A knew this concept, to which they replied “I guess because all the related stuff is green”.

5. Discussion and Conclusion

Many numerical uncertainty management approaches have been applied to student modelling [10]. However these require a more formal network or ontology structure which differs from our light weight approach.

Our current approach is not without limitations. In this paper we only discuss the reasoning about coarse grain concepts in the case where there is no direct evidence. When there are coarse grain concepts with few (say one or two) sources of evidence, the reliability of the concept’s resolved scores is decreased. In future work, we need to consider the amount and type of evidence required by the Standard Student to get a perfect score compared to that of the student.

A second issue is the attributes of the relationship in the ontology. The relationships are (in the case of using Mecureo) not only typed, but also weighted. The formula presented in (2) will have to be modified to take this into account.

Based on the results of the user study, the approach we propose seems promising. The participants understood the interface and they did consider the results of the inference reasonable. The granularity of the concepts was also realized and the participants could appreciate the fact that reasoning was required about higher level concepts that did not have direct evidence sources.
Acknowledgements

We thank Hewlett-Packard for funding parts of this research.

References