

Creating User Models from Web Logs

Judy Kay and Andrew Lum
School of Information Technologies
University of Sydney, NSW, 2006
Australia
{judy, alum}@it.usyd.edu.au

ABSTRACT

This paper describes the ground work we have done in our goal to create an adaptive educational website. We are interested in creating rich user models from a variety of sources. One important source of such user models can be based upon data from the time the user spends at a web page with an audio content. We report basic analyses of such logs from a course whose lectures are delivered as on-line audio associated with web pages.

Keywords

Adaptive systems, online learning, user modeling, visualization, web log analysis

INTRODUCTION

Our goal is to create user models that will allow us to adapt online courses to students' proficiency and learning style. The core of personalization is a model of the user. For example, a personalized teaching system's personalization is driven by its knowledge of the student. An especially important aspect of such user models is the representation of the learner's knowledge since this enables the teaching system to base its teaching on solid foundations of student knowledge. There are many ways to acquire such a model, but one attractive approach is to exploit the trail of evidence of learning as a student makes use of online teaching resources. It is important that the students be able to control and understand their own user models, not only from a privacy standpoint, but also for reflection. There are three main tasks in this research.

- The first is to a model the course that contains the concepts taught and the semantic relationships between them. We have developed two tools, Mecureo [1] and Metasaur [2] to create a lightweight ontology [3] that describes a course.

- The second task is to create rich user models by overlaying the ontology with evidence of student understanding of particular concepts. We envision each time a student accesses a resource that teaches a particular concept, evidence will be added to the user model that will represent an 'understanding' the user has gained from using that resource.
- The third task is the provision of an interface to allow users to scrutinize their user model.

This paper describes our progress in analyzing web log data for evidence of understanding of concepts in the course. There are existing tools available such as those described in [4, 5] for collecting low level user interactions and monitoring student activity. In contrast, the approach we describe in this paper takes a look at web logs from a much higher point of view, examining individual online lectures and the slides within them. The paper begins by describing the teaching context of this work, then the visualisation and the results gathered from the web log data for a course taught at this university. We conclude with a discussion of the results and our plans for future work.

ONLINE TEACHING

The *User Interface Design and Programming* course is taught in the February semester at this university. The course is taught through a combination of online material and face to face tutorials. The course has a website that is customised for each student by presenting material such as pre-recorded lectures and laboratory exercises relevant to their enrolled course (normal, advanced and masters).

The course consists of 20 online lectures that students are expected to attend at times they can choose, but partly dictated by the assignment deadlines. There are recommended deadlines for attending each lecture. Each lecture has around 20 slides, and each slide has associated audio by the author. Generally this audio provides the bulk of the information, with the usual slide providing a framework and some elements of the lecture. This parallels the way many lecturers use overhead slides for live lectures. An example of a slide is shown in Figure 1.

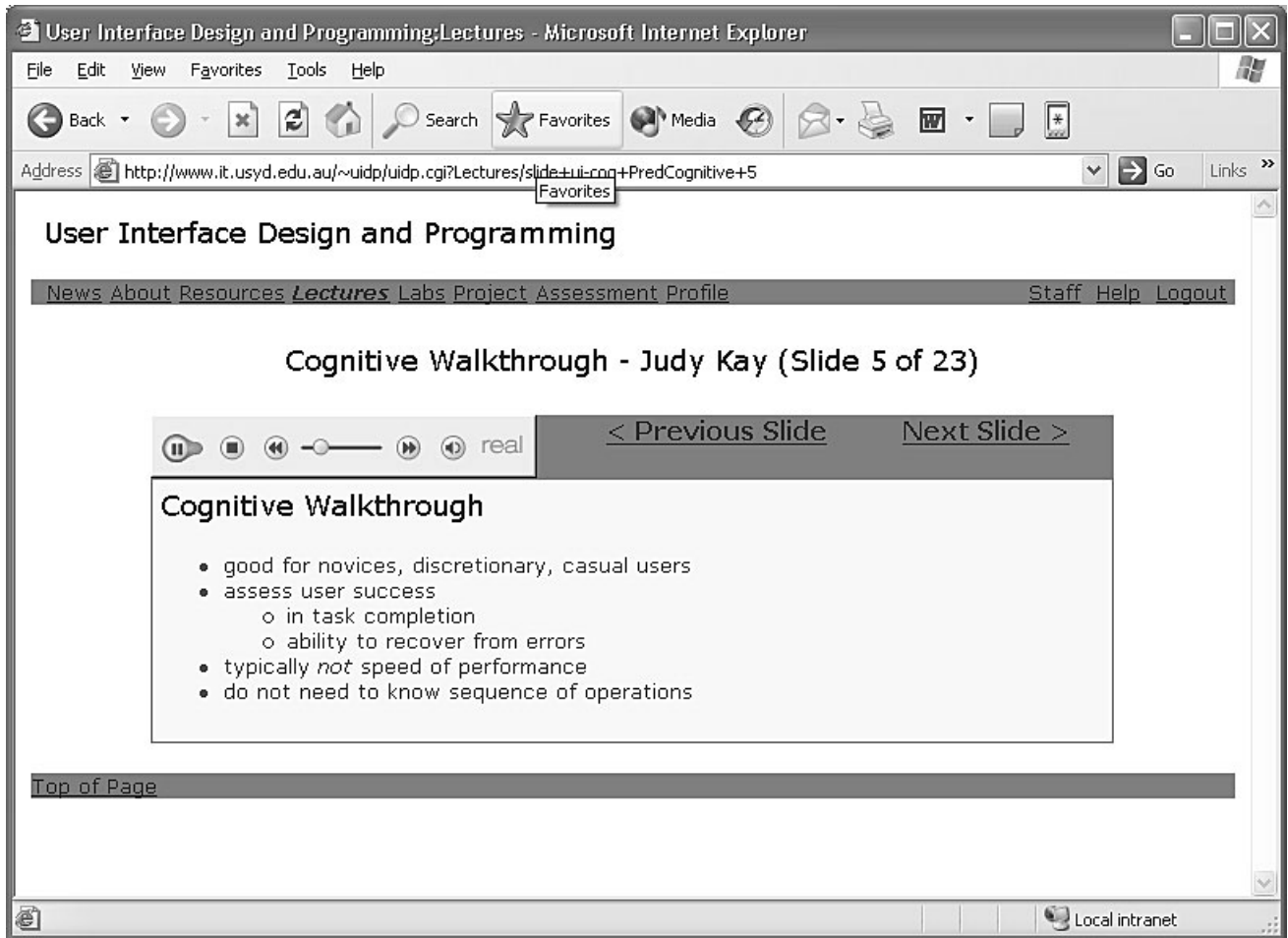


Figure 1. The User Interface Design and Programming course website. This screenshot shows one of the audio slides from the *Cognitive Walkthrough* lecture.

Students should not only view the slides but also listen to the audio (and make their own notes) in order to gain an understanding of the concepts discussed.

A user profile keeps track of student marks and lecture progress. These profiles along with the web access logs can provide user data required to have a good foundation for creating rich user models.

USER MODEL VISUALISATION

Once we have created rich user models, we need to provide an interface that allows users to easily inspect and understand the data about them. The Scrutable Inference Viewer (SIV) is an evolution of VIUM, for Visualisation of Large User Models, a tool that can effectively display large user models in web-based systems [6]. VIUM is inspired by the work by Murtagh on his Automatism Storyteller System [7]. User models are structured as digraphs, as each component has interconnections to other components that are related to it.

SIV extends VIUM to allow the visualisation of ontologies. The visualisation is a Java Applet that can be loaded in a standard web browser with support for Java Swing. It has been designed to ensure that there is sufficient screen real estate available for the display of other, related information so that this can be studied in conjunction with the visualisation.

The concepts in the ontology are displayed in a vertical listing. It utilises perspective distortion to enable users to navigate the user model. At any point in time, the concept with the largest font is the one currently selected. A subgraph is created encompassing this term and those that are deemed related. Concepts connected directly to the selected concept are put into a secondary focus, appearing in a larger font size, spacing and brightness than those further away in the ontology. Similarly, concepts at lower levels in the tree are shown in progressively smaller fonts, less spacing and lower brightness. Concepts that are not relevant are bunched

Table 1. Log Summary for Cognitive Walkthrough Lecture

Slide No.	Actual Time	Seen	Partial Heard	Full Heard	Overheard	Total
1	68	25%	22%	29%	24%	308
2	63	20%	16%	49%	15%	242
3	61	13%	14%	51%	22%	216
4	58	11%	16%	56%	18%	204
5	186	17%	11%	55%	17%	197
6	178	17%	17%	49%	18%	199
7	69	12%	14%	51%	23%	189
8	76	11%	14%	56%	19%	174
9	113	16%	13%	61%	9%	173
10	77	9%	15%	57%	19%	171
11	114	20%	16%	51%	12%	177
12	65	10%	19%	55%	16%	166
13	84	11%	18%	57%	14%	159
14	22	3%	19%	31%	46%	156
15	89	15%	24%	50%	12%	155
16	91	14%	25%	47%	15%	155
17	49	9%	26%	48%	17%	151
18	60	11%	22%	61%	7%	150
19	17	1%	22%	28%	49%	149
20	101	18%	19%	49%	14%	152
21	104	11%	18%	54%	17%	147
22	47	6%	15%	58%	21%	150
23	155	14%	78%	7%	1%	147
Average		13%	21%	48%	18%	

together in a small dimmed font. Users can navigate through the ontology by clicking on a concept to select it. The display changes so that the newly selected concept becomes the focus. A slider allows users to limit the spanning tree algorithm to the selected depth.

The visualisation uses colour and horizontal positioning to show the score and certainty respectively. A value that can be adjusted by the user can be used to alter the colours – scores above the user's chosen value are shown in green, ones below are shown in red. The further away from this value, the stronger the colour hue. The greater the certainty, the more to the left the concept is positioned.

WEB LOG RESULTS

The website for the course collected a form of augmented web log. We have analysed this data to form one source of user modeling evidence. Essentially, if a learner has 'attended' an online lecture, we treat this as evidence supporting the conclusion that they know concepts taught in that lecture. This section describes our analysis of the web logs to model how well each student has 'attended' the lecture.

Table 1 represent an aggregation of the web log data for one lecture from the User Interface Design and Programming course that ran in February semester of 2003. We are interested in how students listened to and reacted to the content on the site. In particular, we are currently examining the length of time users stayed on each slide and whether this contributed to their understanding of the material or not.

We have chosen to aggregate the hits to particular slides in a lecture by the length of time a user has spent on it. Each row of the table corresponds to a single slide in that particular online lecture. The columns are described below:

- *Actual time* is the time of the audio (in seconds)
- *Seen* is the proportion of hits to the slide that stayed for less than 10% of the audio length, as a percentage of the total hits.
- *Partial Heard* is the proportion of hits to the slide that stayed more than 10% but less than 80% of the audio length, as a percentage of the total hits.
- *Full Heard* is the proportion of hits to the slide that stayed more than 80% to 150% of the audio length, as a percentage of the total hits.

- *Overheard* is the proportion of hits to the slide that stayed over 150% of the length of the audio.
- *Total* is the number of hits recorded in the web logs to that slide.

Analysis

We have analysed nine of the online lectures for the course that cover the theoretical aspects of interface design. There are some interesting features of our data. We have been able to identify some trends described in more detail below.

Slides with a short audio time result in most people's visit being recorded as *overheard*. See, for example, Slide 19 in Table 1. To listen to all the audio, a person will need to have either done a *full heard* or an *overheard* of each slide (i.e. listen to all the audio on that slide). The first slide has a high number of hits due to people opening the first slide of the lecture to see what it is about (or to get a visit recorded in their profile).

On current analysis, it seems that the very long slides, with audio over 300 seconds, tend to have relatively low proportions of students 'attending' until the end of the slides. We do not see this effect in the lecture of Table 1. We are currently exploring this in greater depth. Note that the data in Figure 1 is an aggregate of all visits over all users. We are still analysing the individual user data and correlating data about accesses to the lectures slides against performance on examination questions.

All of the online lectures follow a similar trend in the number hits to each slide as the lecture progresses. The first slide has a very high number of hits compared to any other slide in the lecture. The number of hits to the slides then gradually decreases till they reach a stable number (150 hits). This is more than likely due to the fact that the students start the lecture and lose interest and stop early, or are just curious as to what the lecture is about, visiting the first few slides. The profile data on the website also indicated to students whether they have visited the first slide of the lecture or not, and a lot of students will have visited to get that check next to their name to say they have.

One of the striking things to be observed in Table 1, and the other lecture we have analysed, is the generally stable proportions of visits in each of the categories. With the exceptions we have already discussed, the proportions of each visit category duration seem very stable across slides. The values also correlate to the column averages shown at the bottom of Table 1. This seems to reflect individual user's patterns of use of these resources. This is rather odd and we will explore it further in terms of per-user analyses. We will also study the sequence of visits by each user to explore one explanation of our observations - that some users may be visiting each page

of the lecture for a brief viewing and no more. Others may be listening just to a little of each slide.

CONCLUSIONS

We have managed to extract some interesting usage patterns from the web logs. The main challenge has been in massaging the log data into a form we can easily understand and visualize. We have found a consistent pattern in the time students spent on the slides, as well as anomalies with the boundary slides and times. These deviations we have offered our explanations for, although further study and consultation with students who have done the course may yield further insights.

A next stage is to produce individual user models, turning the log data for a user into discrete values that describe their understanding of course concepts. We have made progress into annotating the slides with the Metasaur tool and performed a qualitative evaluation on the validity of the concepts presented in a slide. Each slide will teach a number of concepts (on average about five) and each concept will form a component in the user model whose value will be interpolated from the evidence we have described in this paper.

ACKNOWLEDGMENTS

We thank Hewlett-Packard for funding this research and Bernard Burg for his comments and suggestions.

REFERENCES

- [1] Apted, T. and Kay, J. *Automatic Construction of Learning Ontologies*. In: L. Aroyo and D. Dicheva, Editors. *International Conference on Computers in Education*. Technische Universiteit Eindhoven, p. 55-62. 2002
- [2] Kay, J. and Lum, A., *An ontologically enhanced metadata editor*, TR 541, University of Sydney, Australia.2003.
- [3] Mizoguchi, R., *Ontology-based systemization of functional knowledge*. 2001
- [4] Thomas, R., et al. *Generic usage monitoring of programming students*. In: *ASCILITE 2003*. 2003
- [5] Judd, T. and Kennedy, G., *Extending the Role of Audit Trails: A Modular Approach*. *Journal of Educational Multimedia and Hypermedia*. **10**(4): p. 377-395, 2003
- [6] Uther, J., *On the Visualisation of Large User Model in Web Based Systems*, PhD Thesis, University of Sydney.2001.
- [7] Murtagh, M., *The Automatist Storytelling System: Putting the Editor's Knowledge in Software*, Masters, Massachusetts Institute of Technology.1996.