Evaluating the automatic and manual creation process of adaptive lessons

Maurice Hendrix, Alexandra Cristea and Mike Joy
Department of Computer Science, The University of Warwick, Coventry CV4 7AL, UK
{M.Hendrix, A.I.Cristea, M.S.Joy}@warwick.ac.uk

Abstract

Using adaptive, personalized courses is rewarding, as it can create a better learning experience, tailored for a specific learner's needs. The process of creating these courses, however, is cumbersome and time consuming. Authors can be helped by employing automatization methods that relieve the authoring burden. We previously proposed a method which uses information from an author's Desktop to suggest extensions to existing courses. In this paper we report an experimental evaluation of the method, focusing on usability aspects.

1. Introduction

Authoring Adaptive Hypermedia can generate valuable personalized (learning) experiences [4], but it is known to be a difficult and time-consuming task [5]. A possible solution to this problem is to use automatically generated authoring as much as possible, and there has already been research into how to automatize authoring in different ways [10]. A good basis is to use already annotated resources (such as the Sematic Desktop [7][15]), which can be automatically retrieved when necessary, as dictated by the authoring process. In a Semantic Desktop, resources are categorized by rich ontologies, and semantic links can express various kinds of semantic relationships between the resources. For a file representing a paper, for example, the Semantic Desktop stores not only a filename, but also information about where it was published, when, by whom, etc. These metadata are generated automatically, and stored in the user's personal data store in RDF format. This rich set of metadata makes it easier for the user to semi-automatically retrieve appropriate material for different contexts, for example, when a teacher wants to select materials that fit a certain lecture. In this context, an author has to create some basic lesson material, which then serves as a framework for the final lesson to be created.

In [11] and [12], we previously described the interaction and exchange of data between the adaptive hypermedia authoring environment MOT ("My Online Teacher", pronounced like the French word for "*word*") [8][13], and the Beagle++ environment [1][6].

MOT [8] is a concept-based adaptive educational hypermedia authoring environment with adaptive authoring support. It is an advanced system for authoring personalized e-courses based upon the LAOS [9] authoring framework and offers a web forms interface for the authoring process. The main parts of the LAOS framework it offers are the Goal & Constraints Model and the Domain Model. Elementary blocks of content are represented in the Domain Map, and in the Goal & Constraints Map blocks from domain maps are brought together. This forms an initial version of what end-users (students taking a course) will see before any adaptation is applied to it.

Currently, there are two versions of MOT, the original version, and a new one. In Figure 1 below we see a snapshot of the interface for authoring Domain Maps in the *new* version of MOT.

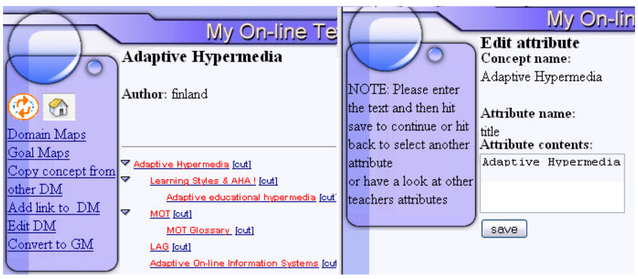

Figure 1 Domain authoring in the new MOT

Beagle++ is an advanced search and indexing engine for the semantic desktop. It is an extension to the Beagle [1] [6] search tool which generates and utilizes metadata information, and keeps a metadata index of all files. Initially extraction tools are used to populate this metadata index.

Our approach uses a standalone Java application called the *Enricher* (or *Sesame2MOT conversion*) to implement the link between the MOT and Beagle++ systems described above. As we can see in the overview below, the Sesame2MOT conversion works by reading current courses from MOT in an XML format called CAF (Common Adaptation Format) and querying the Metadata index kept by Beagle++.

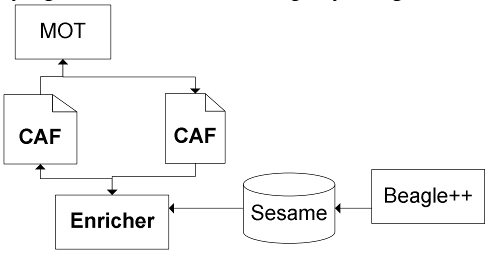

Figure 2 conversion overview

In this paper we evaluate our approach in general, and the prototype [12], which has been constructed, in particular. As MOT has a new version, we evaluate whether this new version is indeed preferable, and whether we should base development of our prototype on this version of MOT.

The remainder of this paper is organized as follows. Section 2 gives a scenario, which illustrates the extraction of learning resources via an automatized process. In Section 3 we explain our evaluation methodology, and in section 4 the results of the evaluation are presented. Finally in section 5 we discuss what these results mean for our approach in general and for the prototype in particular.

2. Illustrating Scenario

We use a scenario for adaptive authoring that builds upon a combination of automatically and manually generated metadata, as introduced in [12].

Prof. Smith is an hypothetical lecturer who is preparing a new course. His university allocates a limited amount of time to this. He uses MOT

- because he considers it useful to be able to extend the course in the future with more alternative paths guided by adaptivity, and
- because he wants to benefit from automatic help during the authoring process.

. This takes slightly more time than static course creation, as the course has to be divided into conceptual entities with explicit, independent semantics and semantic labeling.

The advantage is that the adaptive authoring system can afterwards automatically enrich the course based on pedagogical strategies. For example, the version created by Smith can be considered as the version for beginner students. For advanced students, such as those wishing to pass the course with high marks, the adaptive authoring system can use the Semantic Desktop search to automatically find on Smith's desktop any existing scientific papers that are relevant to the current course. These papers could then be used as alternative or additional material to the main storyline of the static course. This mechanism builds upon the following assumptions.

- Since Smith is a specialist in the subject he is teaching, his interest is wider than that given by the limitations of the course; he therefore both publishes and reads papers of interest on the subject, which are likely to be stored on his computer.
- His papers can be considered as useful extra resources for the current course, and can therefore be reused in this context.
- The storing process has taken place over a long period of time, and Smith may not know exactly where on his computer each article relevant to the current course is.
- Smith has been using the Beagle++ Semantic Desktop System to store both papers and all relevant metadata automatically in RDF format.

This situation can be exploited by the authoring tool; a search will find some of Smith's own papers on the course topic, as well as some papers written by his colleagues on the same topic. He may have saved these papers by himself, received them by e-mail from a colleague, or may have bookmarked them using his browser.

In order for these retrieved resources to be relevant to the course, two conditions have to be fulfilled:

- the domain concept in the course where each resource is most relevant has to be found (the right information), and
- the resource must then be bound to that particular domain concept (in the right place).

How can Smith find the right resource and store it in the right place? The automatic search can take place via the keywords labeling both the course components created by Smith and the matching keywords labeling the papers and resources on his desktop.

How Smith can enrich his course automatically, without much extra work, as well as keep at all times overall control and a coherent overview, is described in more detail in [12]. The following sections evaluate this specific approach, as well as a prototype for it which has been implemented, and discuss possible improvements.

3 Evaluation

The evaluation of this conversion process has taken place in two steps so far. The first step was a small-scale qualitative experiment with 4 PhD students of the IMPDET course organized by the University of Joensuu in Finland, based on the think-aloud method [14]. As can be found in [12], the system was mainly understood, but respondents were unable to provide feedback on the method itself. Some shortcomings of the user interface of the prototype were identified.

The second step was on a larger scale, and contained a great deal of quantitative evaluation, although qualitative information was also sought. In this paper we will focus on the second step and in particular the usability feedback.

3.1 General evaluation setup

This second evaluation was conducted at the Politehnica University of Bucharest in January of 2007, and took place within an intensive two-week course on "Adaptive Hypermedia and The Semantic Web", which was delivered as an alternative track to the regular "Intelligent Systems" course. The students were 4^{th} year undergraduates in Engineering studies and 2^{nd} year Masters students in Computer Science, all from the English-language stream. Firstly, basic knowledge on Adaptive Hypermedia and Semantic Web was addressed – the first course week was dedicated to theory, and finished with a theoretical exam. Out of the initial 61 students, only the students with a satisfactory knowledge of the theory were selected to continue with the practical part. The 33 students that passed the theory exam worked with the two versions of MOT (old versus new) and the Sesame2MOT (Enricher) conversion, the prototype constructed for the automatic authoring approach [11]. After these experiments, they were requested to submit questionnaires, to answer both generic and specific issues regarding the automatic generation of adaptivity and personalization. The questionnaires consisted of five parts; first a SUS [3] questionnaire for each of the three systems, and then two more specific questionnaires, for the Sesame2MOT conversion and for the comparison of the new version of MOT with the previous version. Here we mainly focus on the usability aspect of the evaluation process.

3.2 Hypotheses

We based our evaluation on the following generic hypotheses.

1. The respondents have acquired more knowledge than they initially had with the help of the theoretical course (explanation) part.
2. The respondents enjoyed working as authors in the three systems from a usability perspective.
3. The respondents' overall preference (from a usability perspective) is as follows, in increasing order: old MOT, new MOT, Sesame2MOT.
4. The new MOT is more usable than the old version; hence we should base further developments on this version of MOT.

In addition to these four generic hypotheses, we identified a number of more specific hypotheses which we tested using the data we collected. These hypotheses were constructed from the main hypotheses. For example from hypothesis 2 the following sub-hypotheses, were derived.

- The respondents enjoyed working with the OLD MOT.
- The respondents enjoyed working with the NEW MOT.
- The respondents enjoyed working with the Sesame2MOT conversion.

To confirm all the sub-hypotheses, respondents were requested to fill in a SUS questionnaire for the OLD MOT, the NEW MOT and the Sesame2MOT conversion. At the very beginning a pre-test was held. This pre-test was identical to the first part of the theoretical exam respondents had to take. Therefore, by comparing the results of the pre-test and post-test we were able to compare the knowledge respondents had on the subject before and after the course.

4 Evaluation Results and Discussion

For testing our hypotheses we used numerical averages, and tested their significances with the help of a T-test. We assumed a confidence of 95% would be reasonable. The T-test establishes tests whether the difference between a value and the average of a sample or the averages of two samples is significant. For a hypothesis to be supported confirmed the difference needs to be significant and be in the direction the hypothesis suggests. For example is we test the difference between pre-test and post-test exam and have as hypothesis that respondents did better in the post exam, the average from the post exam must be higher and the difference between pre- and post exam needs to be significant. The difference between two samples or a sample and a value is considered significant if the probability P that the difference arose by chance is $P<0.05$. Due to lack of space, we shall focus on the usability of the different systems, in order to highlight further development of the Sesame2MOT converter.

4.1 SUS

With a SUS score the usability of systems can be compared. The average score can be contrasted and visual graphs can be constructed to identify specific problem points. The questions (which are alternately positive and negative) are plotted on a circle using a scale from 1 (strongly disagree) to 5(strongly agree), with 1 in the centre and 5 at the border. If the results for the questions are placed on the scale, the ideal system should show a perfect star shape, as positive and negative questions alternate. In Figure 3 below, the SUS scores for the different systems are shown in such a SUS graph. The figure shows that the systems have relatively similar scores. Visible differences are that Sesame2MOT seems to have a higher perceived learning threshold, whereas the old MOT is considered more inconsistent and more cumbersome.

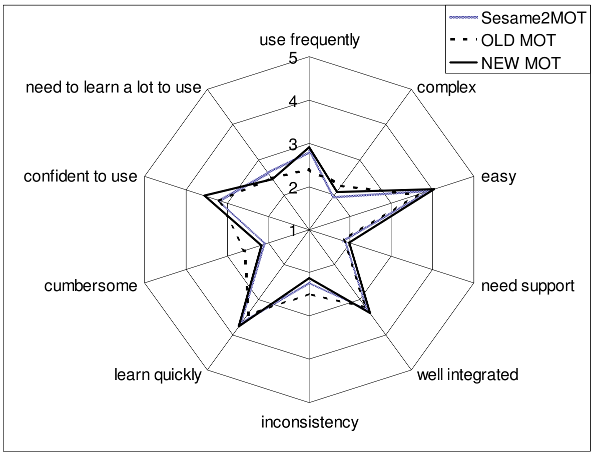

Figure 3 SUS score for the three systems

Normalized responses range from 0 to 4, see [3]. Thus we applied a T-test comparing the normalized results against the average neutral value of 2. A paired T-test was used, since we compared answers of the same sample (group of students). Moreover, the main hypotheses were further broken into sub-hypotheses.

4.2 General Hypotheses

Below we list the main hypotheses and comment on how much they are supported.

1. *The respondents have acquired more knowledge than they initially had with the help of the theoretical course (explanation) part.*
   Comparison of the grades obtained from the pre-test and post-test shows that the post-course test grades are significantly higher (difference 5.75 out of 10; p=0.00<0.05; t=25.59), hence the hypothesis is *supported*.

2. *The respondents enjoyed working as authors in the three systems from a usability perspective.*
   The results for the old MOT (mean 2.39 (expected >2); p=0.519>0.05; t=0.65) and Sesame2MOT (mean 2.78 (exp >2); p=0.095>0.05; t=1.72) on enjoyment were not significant. The respondents did significantly enjoy working with the new MOT (mean mean 2.97 (exp >2); p=0.01<0.05; t=2.66) The hypothesis as a whole *cannot be supported*. This is possibly due to the formal setting of the course.

3. *The respondents' overall preference, from a usability perspective, is as follows, in increasing order: old MOT, new MOT, Sesame2MOT.*
   The results on learning preferences, and the preference for Sesame2MOT over the new MOT (difference -0.07 (>0 exp.); p=0.18<0.05; t=-1.44) were not significant. The hypothesis *cannot be supported*. Preference for the new over the old MOT (diff 0.26 (>0 expected); p=0.00<0.05; t=4.16) was confirmed.

4. *The new MOT is more usable, hence we should base further developments on this version of MOT.*
   For all different parts, as well as overall SUS score (see hypothesis 3), the new version of MOT is preferred over the old version. Thus we should indeed focus further development on the new version. The hypothesis *is supported*.

4.3 Further results

The SUS questionnaire covers more issues than the focus of the current paper.
For instance, none of the hypotheses related to learning threshold showed any significant difference between the three systems. This is possibly due to the fact that systems respondents had to learn all the theory before working with the three systems, or that both MOTs are very similar from a theoretical point of view.
We computed the correlation between the SUS scores for the 3 different systems. This showed that the respondents' answers to all three systems' SUS questionnaires are significantly correlated. This seems to be due to one of the following two reasons:

- respondents were not quite aware for which systems they were filling in the SUS questionnaire (suspicion based on some questions from students)
- or the students perceived the three systems as variants or parts of the same system.

Moreover, we also found that the correlation between the scores for the new MOT and for the Sesame2MOT conversion is highest. This could indicate that a substantial number of respondents viewed the Sesame2MOT conversion and the new MOT as one system, since Sesame2MOT is currently integrated into the new MOT.

5 Conclusions

In this paper we have briefly reviewed an authoring environment for personalized courses, as well as an Enricher mechanism and prototype based on Semantic Desktop technology. The paper reports the evaluation of this prototype from a usability point of view. We have seen that respondents acquired more knowledge on the subjects involved then they initially had. They liked working with the new MOT and the Sesame2MOT conversion, but were less positive about the old MOT. We cannot confirm the preference order we expected, possibly due to respondents not really making the distinction between the new MOT and the Sesame2MOT systems (as shown by the correlation results). As we have seen, the new MOT is preferred over the old version and we should base development on this version. A highlighted drawback is that Sesame2MOT has a steeper learning curve. We also see that integration with the MOT systems could be improved. From the quantitative feedback gained we suspect that the latter is one of the main causes of the steeper learning curve, and hence future work will be to further integrate the conversion into the new MOT system.

Acknowledgments. We would like to thank the Socrates mobility exchange programs at The University of Warwick, Politehnica University of Bucharest in general, and Prof. Dr. Paul Cristea in particular for giving us the opportunity to conduct our user study.

6. References

[1] Beagle website, viewed 14 February 2007, http://beagle-project.org/Main_Page

[2] Beagle++ website, viewed 14 February 2007, http://beagle.kbs.uni-hannover.de/

[3] Brooke, J.: SUS: a "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A. and McClelland, A.L. (eds.). Usability Evaluation in Industry. London: Taylor and Francis. , 1996

[4] Brusilovsky, P., Adaptive hypermedia, User Modelling and User Adapted Interaction, Ten Year Anniversary Issue, 2001 (Alfred Kobsa, ed.) 11 (1/2), 87-110.

[5] Celik, I., Stewart, C. and Ashman, H. Interoperability as an Aid to Authoring: Accessing User Models in Multiple AEH Systems, Proceedings of A3H: 1st International Workshop on Authoring of Adaptive and Adaptable Hypermedia, 2006

[6] Chirita, P.-A., Costache, S., Nejdl, W. and Paiu, R. 2006. Beagle++: Semantically Enhanced Searching and Ranking on the Desktop. Proceedings of the 3rd European Semantic Web Conference, Budva, Montenegro, 2006.

[7] Chirita, P.-A., Gavriloaie, R., Ghita, S., Nejdl, W. and Paiu, R., 2005, Activity-Based Metadata for Semantic Desktop Search. Proceedings of the 2nd European Semantic Web Conference, Heraklion, Crete.

[8] Cristea, A.I. and De Mooij, A., 2003, Adaptive Course Authoring: My Online Teacher. Proceedings of ICT'03, Papeete, French Polynesia.

[9] Cristea, A.I. and De Mooij, A., 2003, LAOS: Layered WWW AHS Authoring Model and their corresponding Algebraic Operators. WWW03 (The Twelfth International World Wide Web Conference), Alternate Track on Education,Budapest,Hungary, 2003.

[10] Cristea, A.I. and Stewart, C., Automatic Authoring of Adaptive Educational Hypermedia, book chapter in "Web-Based Intelligent e-Learning Systems: Technologies and Applications", Zongmin Ma. IDEA Publishing Group, 2005

[11] Hendrix, M., Cristea, A.I., Nejdl, W., Authoring Adaptive Learning Material on the Semantic Desktop, 1st International Workshop on Authoring of Adaptive and Adaptable Hypermedia, Adaptive Hypermedia Dublin Ireland, 2006

[12] Hendrix, M., Cristea, A.I., Nejdl, W., Authoring Adaptive Educational Hypermedia on the Semantic Desktop, to appear in International Journal of Learning Technologies (IJLT), 2007

[13] MOT homepage, viewed 14 February 2007, http://prolearn.dcs.warwick.ac.uk/mot.html

[14] Nielsen, J., Usability Engineering. Boston: Academic Press, 1993

[15] Semantic Desktop, viewed 14 February 2007, http://www.semanticdesktop.org/