In Paul De Bra (Ed.) (2002). *Adaptive Hypermedia and Adaptive Web-Based Systems: Proceedings of AH 2002*. Berlin: Springer.

Pros and Cons of Controllability: An Empirical Study

Anthony Jameson[1] and Eric Schwarzkopf[2⋆]

[1] DFKI / International University in Germany
[2] DFKI, the German Research Center for Artificial Intelligence

Abstract. A key usability issue with systems that adapt to their users is *controllability*: the ability of the user to determine the nature and timing of the adaptation. This paper presents an empirical study of the trade-offs involved in an attempt to ensure a suitable degree of controllability. Within an adaptive hotlist for a conference web site, two mechanisms for providing users with recommendations of conference events were compared: *automatic* vs. *controlled updating* of recommendations. In an experimental setting, each of 18 users worked with both variants of the adaptive hotlist, as well as with a nonadaptive variant. The users differed markedly in their responses to automatic vs. controlled updating. A number of reasons for these differences could be found in the objective and subjective data yielded by the study. The study illustrates how preferences for different forms of user control can be influenced by factors ranging from stable individual differences to unpredictable features of a situation. General implications for the design of controllable adaptive systems are discussed.

1 Is Maximal Controllability Always Best?

One of the main usability issues in connection with systems that adapt to their users concerns *controllability*. For example, Norman [5] wrote in an influential article: "An important psychological aspect of people's comfort with their activities—all of their activities, from social relations, to jobs, to their interaction with technology—is the feeling of control they have over these activities and their personal lives" (p. 69). The relationships between controllability and other usability issues with user-adaptive systems have been discussed by Wexelblat and Maes [8] and by Jameson [3, section 4].

One plausible policy for designers of user-adaptive systems is to give users maximal control over all aspects of system adaptation. For example, Kay [4] discusses many ways in which learners can be given control over learner-adapted teaching systems. But as Kay also points out (p. 121), simply maximizing control in all respects may not always be the best policy: Some users may have generally less desire for control than others; and making the user exercise too much control may lead to distraction and time-wasting.

⋆ The research described here is being supported by the German Ministry of Education and Research (BMB+F) under grant 01 IW 001 (project MIAU). Björn Mittelsdorf and Marie Norlien made essential contributions to the empirical study.

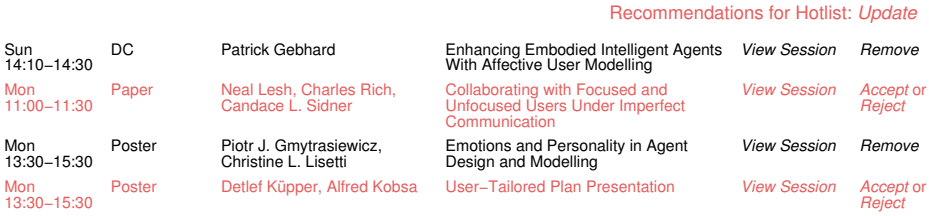

Fig. 1. *Example hotlist from the UM 2001 web site.*

(The first and third entries were added by the user; the second and fourth, shown in the system in a red font, are recommendations made by the system. Italicized words represent hyperlinks, referred to in this article as "buttons".)

Similarly, Trewin [7] discusses the controllability trade-offs involved with agents that help users to configure aspects of an operating system or an input device such as a keyboard. For example, a physically impaired user may not be able to operate a keyboard even well enough to initiate and control the configuration process; so fully autonomous system adaptation may actually give the user more control overall than if she were required to control the configuration process.

Although there has been much discussion among researchers about controllability, some of it quite heated, there is a dearth of systematically gathered evidence about what users themselves think about these issues. The present study aims to provide such evidence within the context of one particular adaptive hypermedia system. Section 2 introduces the system, Section 3 describes our empirical study, and Section 4 presents and discusses the results.

2 The Hotlist And Its Recommendation Component

The Eighth International Conference on User Modeling, UM 2001, held in July of 2001, was the latest in a biennial series of conferences concerning user-adaptive systems (see [1]). It offered the following *conference events*: 3 invited talks, 3 tutorials, 19 full paper presentations, 21 poster presentations, and 12 doctoral consortium presentations. The conference web site (http://dfki.de/um2001) introduced a variety of adaptive features, of which only the *hotlist* will be examined in the present study. The hotlist (see Fig.1) is basically a specialized bookmarking tool that helps a potential attendee put together a list of personally relevant conference events. Once the user has explicitly added some events to the hotlist, the system can insert a set of *recommendations*—essentially, a set of similar events that this user might be interested in.

The recommendations are computed with a naive Bayes classifier (see, e.g., [6]), using as features a set of 22 domain-specific key concepts such as *Machine Learning*. If the user clicks on the *View Session* button for a recommended event to see its full description, she will also see a simple "explanation" of the recommendation in terms of the system's estimates of the user's interest in the individual key concepts associated with the event. Further details concerning the recommendation mechanism must be

omitted here for reasons of space; they are not required for an understanding of the results that will be presented below.

After evaluating a recommendation, the user can choose to *Accept* the recommendation, making it into a normal hotlist entry; or to *Reject* it, causing it to disappear from the hotlist.

At various times, the system *updates* the set of recommendations: It removes any recommendations currently in the hotlist and replaces them with a (perhaps overlapping) set that is based on all of the user's relevant actions so far. Different ways of controlling this updating process were compared in our empirical study.

3 Empirical Study: Issues and Method

It would be possible to design a study to determine what type of control was really best for users in the long run. But it is equally interesting to find out how users deal with and respond to each system variant during an initial encounter of just a few minutes. After all, users often briefly try out a system—or an option within a system—and decide on the basis of a small sample of experience whether to continue using it.

Moreover, the goal of the empirical study is not to determine the accuracy or overall utility of the hotlist recommendations. Instead, it is assumed (as will be confirmed) that the recommendations have only modest accuracy, as is the case with many recommender systems, because of the severely limited evidence on which they are based. The question is: How much control do users want to have when dealing with these imperfect recommendations?

Subjects

Subjects were 17 students and 1 recent graduate from Saarland University and the International University in Germany. Only subjects were recruited whose major or minor course of study had some affinity with the topic of user modeling (e.g., computer science, information science, or psychology), so that the experimental task (to be described below) would be motivating and manageable to them; but the large majority had little or no specific knowledge of the field. The number hours per week that subjects reported spending in the world-wide web averaged 12.9, with a standard deviation of 10.8. All subjects were male. They received 15 German marks for their participation.

System Variants Studied

Three variants of the hotlist were used:

- *Controlled updating of recommendations.* This is the variant shown in Fig. 1: The user explicitly requests each update of the recommendations by clicking on the *Update* button at the upper right.
- *Automatic updating of recommendations.* In this variant there is no *Update* button; the system updates the recommendations automatically whenever the user adds or removes a hotlist event or accepts or rejects a recommendation.

Instructions (paraphrase):

You are working as a research assistant at the nearby research institute [name given]. A number of more senior researchers at this institute are considering attending the UM 2001 conference, which will take place 3 months from now.

To spare these researchers the time of familiarizing themselves with the conference site and program, the director has asked each of them to send you an email message in which they characterize the topics that they are interested in.

For each such message, your job is to build up a list of relevant conference events, which you will email back to the researcher in question. On the basis of this list, the researcher will decide whether he or she considers it worthwhile to attend the conference.

Example email message:

From: Anna Reiter <[local email address]>

Subject: What I'd like to see at UM 2001

For me, the most important methodological approach in the area of user modeling is machine learning. Often, methods from this category are applied in web-based systems, or in systems that select specific news stories for individual users. I'm *not* interested in these last two types of application of machine learning.

Anything that deals specifically with the improvement of automobile safety would be especially interesting to me.

Another thing I'm interested in is systems that model some type of psychological state of the user, such as emotions or stress.

Best regards, Anna Reiter

Fig. 2. *Paraphrase of the key instructions (left) and one of the three fictitious email messages used as a sketch of an interest profile (right).*

- *No recommendations*: In this variant, the user can use only the basic hotlist, adding or removing events but receiving no recommendations.

In a within-subject design, each subject used all three variants of the hotlist, the order of use being counterbalanced as is described below. This type of design was chosen over a between-subject design because of (a) our expectation (confirmed during the study) that individual differences would be very large; and (b) our desire to hear the comparative comments of subjects who had experienced all three variants. Learning effects could not be avoided with this design. But the counterbalancing measures described below ensured that such learning effects could not lead to overall differences in the results for the three variants; and we will also see that the observed differences among subjects are not explainable in terms of learning effects.

Each subject spent only a limited amount of time with each variant: about 4 minutes of introduction plus 7 minutes of measured use. A serious conference visitor might spend considerably more time constructing a personal conference schedule. On the other hand, the shorter amount of time seems typical of the time that a user might spend trying out the hotlist recommendations before deciding whether to continue using them to create a complete schedule.

Material

The experimental task assigned to the subjects was designed to overcome two obstacles:

1. Subjects have considerably less familiarity with the topic of the conference than a potential conference visitor would typically have.
2. Because of the within-subject design, each subject has to search the conference site with respect to three different configurations of interest.

The left-hand side of Fig.2 summarizes the way in which the experimental task was introduced to each subject; the right-hand side of the figure shows one of the three

fictitious email messages employed. Each of the three messages had a similar style and structure, and it described interests for which it was approximately equally easy to find relevant conference events. The interests expressed were in part strongly related to the hotlist recommender concepts, but for the most part subjects had to look at the detailed information about an event in order to decide whether it was really relevant. This situation appears to be typical of the way in which real potential conference visitors use the hotlist.

Orders of Presentation

Each of the 6 possible orders of the 3 system variants was employed equally often (i.e., for 3 of the 18 subjects). Each of the 3 fictitious interest profiles was used equally often in the 1st, 2nd, and 3rd temporal position and equally often together with each system variant.

Procedure

Each subject participated individually with the guidance of an experimenter. In an introductory phase that lasted between 20 and 25 minutes, the experimenter explained that the investigators had developed various methods for searching for information in a conference web site and that they were interested in evaluating and improving them with a view to possible use in other sites. The experimenter then summarized some basic ideas of the field of user modeling and explained the fictitious situation. Using an example email, the experimenter gave an explanation of the web site and the hotlist, frequently stopping to allow the subject to try out the system's functions.

In each of the three main trials, the subject first read one of the emails from a hypothetical colleague and then was allowed 7 minutes to build up a hotlist for that colleague, starting with the system initialized for a new user (with an empty hotlist). In the system's log files, a record was kept of all pages visited and all actions taken in relation to the hotlist. The experimenter took notes on other observable aspects of the subject's behavior. At the end of the 7 minutes, the experimenter saved the hotlist to disk in its printable form.

After the three main trials, the subject typed in answers to a number of questions about his use of the system, some of which are discussed below. Finally, subjects were asked for further comments during a debriefing.

Despite their lack of knowledge about user modeling, subjects reported no major difficulties in understanding the fictitious interest profiles or in evaluating individual events with regard to these profiles.

4 Results

4.1 Quantitative Objective Results

Although the key variable of interest is subjects' subjective evaluation of the two types of updating, some objective results will give us a general picture of the way in which they used the hotlist.

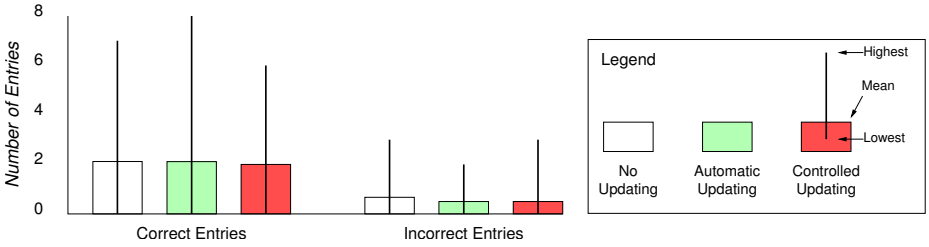

Fig. 3. *Mean numbers of correct and incorrect entries in the final hotlists produced by subjects with the three system variants.*

(The maximum numbers of correct entries for the three interest profiles were 12, 14, and 18 respectively, but some correct entries were difficult to identify as such. The upper and lower ends of the vertical line segment in the middle of each bar indicate the highest and lowest values, respectively, that were found among the 18 values obtained for the 18 subjects.)

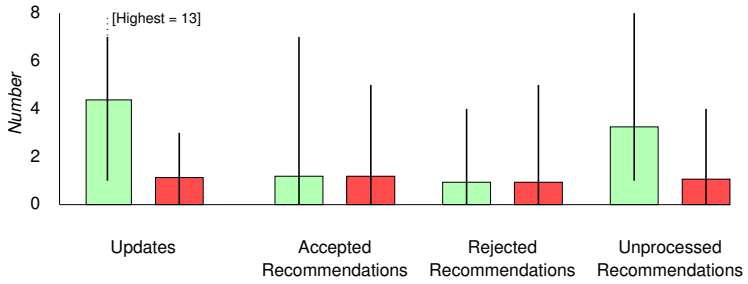

Fig. 4. *Objective results concerning the appearance and processing of recommendations with automatic and controlled updating.*

(Legend as for Figure 4.)

Figure 3 shows that subjects were just about equally (un)successful with all three variants in finding relevant events to add to the hotlist. The rather small overall number of events found is consistent with the limited amount of time that subjects had available to process each interest profile.

Figure 4 shows the differences between the two system variants that included recommendations, in terms of how the system presented recommendations and how users responded to them. It is not surprising that automatic updating led to about 4 times as many updates of the hotlist, since it involves utilizing just about every opportunity for an update.

Although subjects using controlled updating experienced only 30% as many hotlist updates as those using automatic updating, they received 59% as many recommendations: By the time they had gotten around to requesting an update, it was likely to contain more new recommendations than a typical automatic update.

On the average, subjects accepted and rejected exactly the same number of recommendations using the two variants. The big difference is that with automatic updating, many more recommendations were never responded to explicitly at all (this difference,

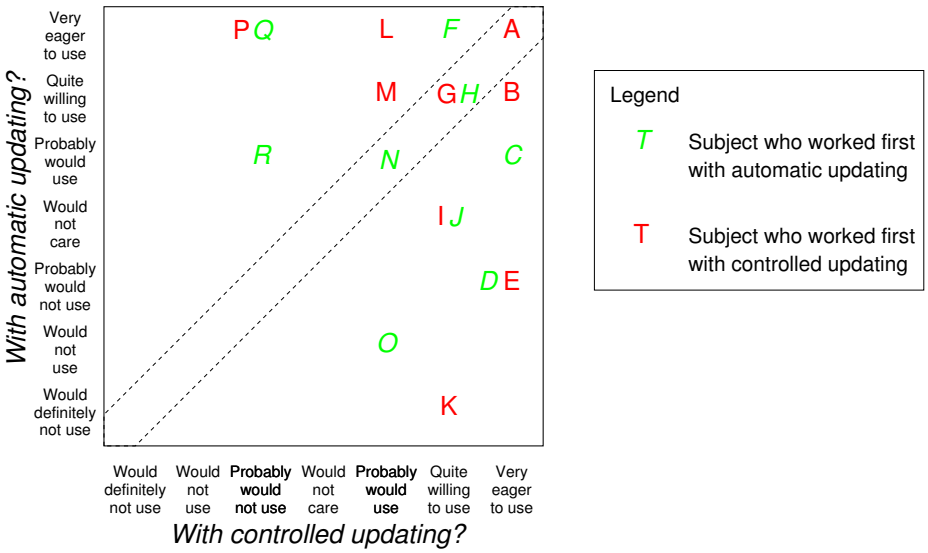

Fig. 5. *Scattergram of subjects' expressed preferences for two updating methods.*

(Each letter represents the responses of one subject. With respect to each method, the question was: "If you had to perform more searches, what would you prefer to do: use the recommendations (that is, have them displayed, and follow up on at least some of them); or not use the recommendations (that is, turn them off or not request them in the first place)?".)

shown in the right-hand pair of bars in Fig.4, is highly significant by a Wilcoxon rank-sum test: $Z = -3.03$, $p < 0.01$). As the logs confirm, in many cases these recommendations were swept away by an automatic update after the subject had made some change to another aspect of the hotlist.

4.2 Attitudes Toward the System Variants

Figure 5 gives an overview of the subjects' responses to two questions that were designed to reveal (indirectly) their preferences for controlled vs. automatic updating. Given the emphasis in previous literature on the importance of controllability, we might expect to see a statistically significant tendency for subjects to prefer controlled updating. (With a within-subject design involving 18 subjects, a moderately strong tendency could have been detected.) Instead, the most important conclusion to be drawn from Fig.5 is that users responded in very different ways to the questions. Given a sufficiently large sample of subjects, we could no doubt find a statistically significant preference for one type of adaptation or the other. But it is more important to understand the reasons for the differences in users' responses, using all of the available types of data: their responses to the rating scales in the questionnaire; the verbal comments that they typed into the questionnaire and made during their work or during the debriefing; and the detailed records of the interaction that can be found in the system logs. This type of analysis inevitably involves qualitative interpretation.

Subjects Who Preferred Controlled Updating

Subject K (cf. Fig.5) is typical of users who have a strong general desire to remain in control of the interaction with a system. He wrote "I am used to updating information manually" and "I hate having the information appear automatically". K's behavior, as it is revealed by the log files, is consistent with his attitude: With controlled updating, he requested 1 update of the recommendations and proceeded to accept or reject each of the 3 recommendations that appeared. With automatic updating, he received 8 recommendations in 4 updates, and he was able to follow up on only 4 of them (1 per update).

Subject A—the most successful subject of all in terms of the number of relevant events found—showed an attitude and a strategy similar to that of K with controlled updating. But unlike K, he was able to follow up on the recommendations equally thoroughly in the automatic updating condition, accepting or rejecting 11 out of the 12 presented—simply because the system happened to present only about 1 new recommendation after each update. Consistent with this result, A expressed an equally strong willingness to work with both system variants. He mentioned two advantages of automatic updating that will be discussed below, and he stated that his true preference would be to switch back and forth at will between the two variants.

Subject O had quite a different reason for preferring controlled updating: On the whole he found the recommendations to be of little value, accepting only 1 of the total of 5 that he received. Accordingly, his attitude toward both of the variants with recommendations was relatively negative (cf. Fig.5). But he was especially critical of the variant with automatic updating, saying that the burden of having to read the recommendations may be even greater than that of reading through the detailed event descriptions. Note that a reasonable strategy is for the user to start paying attention to the recommendations only when the user has reason to believe that the system's model has achieved a reasonable level of accuracy. In both system variants, the user can indeed always decide whether to follow up on the recommendations; but with automatic updating, the user pays a price for the recommendations even when he or she is ignoring them, in terms of screen clutter and longer system response times.

Subjects Who Preferred Automatic Updating

The clearest preference for automatic updating was shown by subject P (see Fig.5). He volunteered the comment that "If you are not accustomed to press the update button periodically or after a decision you just made, you'll miss topics."

Similarly, Subject L commented spontaneously on two advantages of automatic updating: First, he found it "too time-consuming to press the button each time". Second, L appreciated the fact that the automatically generated recommendations always represented the system's most up-to-date model of his interests.

Subject Q illustrated a somewhat different drawback of controlled updating: The danger that the user may forget about updating entirely. Indeed, while using the variant with controlled updating he had completely forgotten about recommendations, using instead just the basic hotlist, as he himself noticed later.

Table 1. *Summary of the potential advantages of each variant of the hotlist recommender that came to light in the empirical study.*

Potential advantage	Precondition(s) for advantage to apply
Controlled updating:	
1. The user's feeling of control over the interaction with the system is enhanced.	The user has a general desire to control interactions.
2. The user can follow up on more than one recommendation in a given set.	The user receives relatively large, nonoverlapping sets of recommendations. The user pursues the strategy of looking at all of the recommendations in each set.
3. System response times can be faster because of less frequent updating.	Technical conditions make system response time an important factor. The user would not choose to request an update at every opportunity.
4. The user can restrict updates to situations in which the system's model of her interests is assumed to have useful accuracy.	The user can assess the likely accuracy of the system's user model.
5. A smaller amount of irrelevant text appears in the hotlist.	The user finds recommendations distracting although they are clearly distinguishable from normal hotlist entries – perhaps because of limited available screen space.
Automatic updating:	
1. The user is regularly reminded that new recommendations are available.	The user's strategy does not provide for regular consideration of the recommendations. The user has not yet learned that hotlist actions typically result in new recommendations.
2. The user is spared the effort of clicking on a button to obtain new recommendations.	The user's hotlist–related actions are sufficiently numerous that new recommendations are frequently available.
3. The recommendations displayed always reflect the system's most complete model of the user's interests.	The accuracy of the system's user model tends to improve significantly with each modification to the hotlist.
4. The user cannot overlook the availability of the recommendation feature.	The user is not yet accustomed to using recommendations.

5 Discussion

Whenever a choice between controlled and automatic adaptation arises, each solution is likely to have its own potential advantages over the other one. The specific potential advantages of automatic and controlled updating that emerged from our study are summarized in Table 1.

As this table illustrates, the relative importance of each of these advantages may depend on various types of conditions:

1. The nature of the application and of the adaptation involved.
2. Individual differences among users in terms of preferences, experience, and ways of approaching the tasks in question.
3. Relatively stable contextual factors such as the speed of an internet connection.
4. Essentially random situational factors such as the nature of the information retrieved during a small number of search attempts.

One general design implication is that an attempt to deal with the controllability problem should begin with an analysis of the reasonably stable, predictable conditions

that are likely to be relevant. For example, Trewin [7] discusses different controllability mechanisms that are appropriate for different types of configuration task.

A second approach to providing suitable controllability is to allow users to choose the type of control that they desire (see, e.g., [8, "Issue 4"]). For example, if our hotlist included a button for toggling between automatic and controlled updating, those users who had a clear, strong preference for one type of updating might be quickly satisfied. But a user cannot in general be expected to be able or willing to take into account all of the relevant considerations (e.g., the entire set listed in Table 1).

To a certain extent, the factors identified as relevant can be taken into account by the system itself. For example, our hotlist recommender could compute at any moment the expected utility of an automatic update, taking into account factors such as the length of the delay that would be caused by the update and the number of recommendations in the hotlist that the user has not yet processed. The user could then be allowed to set an expected utility threshold that must be exceeded before an automatic update is performed. (A similar approach was realized in the LUMIÈRE prototype; cf. [2].)

Given the nature of the factors that tend to be involved, neither the designer nor the user nor the system—nor all of them working together—will in general be able to ensure that the right degree of controllability is available all of the time. It should be anticipated that frustrations like those experienced by our subjects with respect to both of the adaptive variants will in some cases occur; and the possibility should be taken into account that they may cause a user to abandon a system entirely.

Although this last point sounds discouraging, taking into account the limited predictability of users' behavior and responses may be an important step toward an adequate solution of the problem of giving users appropriate control over adaptation.

References

1. Mathias Bauer, Piotr Gmytrasiewicz, and Julita Vassileva, editors. *UM2001, User Modeling: Proceedings of the Eighth International Conference*. Springer, Berlin, 2001.
2. Eric Horvitz, Jack Breese, David Heckerman, David Hovel, and Koos Rommelse. The Lumière project: Bayesian user modeling for inferring the goals and needs of software users. In Gregory F. Cooper and Serafin Moral, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference*, pages 256–265. Morgan Kaufmann, San Francisco, 1998.
3. Anthony Jameson. Adaptive interfaces and agents. In Julie A. Jacko and Andrew Sears, editors, *Handbook of Human-Computer Interaction in Interactive Systems*. Erlbaum, Mahwah, NJ, 2002. In press.
4. Judy Kay. Learner control. *User Modeling and User-Adapted Interaction*, 11:111–127, 2001.
5. Donald A. Norman. How might people interact with agents? *Communications of the ACM*, 37(7):68–71, 1994.
6. Michael Pazzani and Daniel Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27:313–331, 1997.
7. Shari Trewin. Configuration agents, control and privacy. In *Proceedings of the ACM Conference on Universal Usability*, pages 9–16, Arlington, Virginia, U.S., 2000.
8. Alan Wexelblat and Pattie Maes. Issues for software agent UI. Unpublished manuscript, available from http://wex.www.media.mit.edu/people/wex/, 1997.