An Empirical Assessment of Adaptation Techniques

Theophanis Tsandilas
Department of Computer Science
University of Toronto
fanis@dgp.toronto.edu

m.c. schraefel
School of Electronics and Computer Science
University of Southampton
mc@ecs.soton.ac.uk

ABSTRACT
The effectiveness of adaptive user interfaces highly depends on the how accurately adaptation satisfies the needs of users. This paper presents an empirical study that examined two adaptation techniques applied on lists of textual selections. The study measured user performance controlling the accuracy of the suggestions made by the adaptive user interface. The results indicate that different adaptation techniques bare different costs and gains, which are affected by the accuracy of adaptation.

Categories and Subject Descriptors: H5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Interaction Styles, Evaluation/Methodology*.

General Terms: Design, Experimentation, Human Factors

Keywords: Adaptive interfaces, adaptation techniques, user study

INTRODUCTION
Adaptive interfaces have been proposed as solutions for a multitude of sins such as reducing bloat in high-functionality applications [4] and "personalizing" navigation through information spaces [2]. A limitation of automated adaptation is that its success highly depends on the ability of an inference mechanism to accurately capture the needs of users. Previous usability studies [3, 5, 6, 8] have focused on evaluating the overall performance of an adaptive interface in comparison to the performance of its non-adaptive version, disregarding the fact that this performance depends on the effectiveness of the underlying intelligent system. As Tiernan et al. [9] showed, the reliability of the assistance provided by an intelligent system can even affect how users trust and use the system.

This paper presents the results of an empirical study on two adaptation techniques that help users to locate information in lists of textual items. The study examined the techniques in isolation from any particular adaptation mechanism, i.e., we employed a controlled simulation of an inference mechanism. This allowed us to control the accuracy of the suggestions made by the adaptive user interface and examine user performance with respect to this accuracy. Our approach provides a first step towards assessing costs and gains associated with an adaptive user interface in separation from the underlying intelligent mechanisms.

In the following sections, we present the method and the results of our experiment. We conclude with a discussion of our results.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

EXPERIMENTAL METHOD

Goals and Techniques
The main goal of our study was to investigate how the effectiveness of an adaptation technique would change with respect to the accuracy of suggestions made by an adaptive user interface. We tested a specific user interface: lists of textual selections. We focused on goal-oriented tasks as opposed to free-browsing tasks, which can also be supported by lists, but their simulation in controlled experiments is difficult. We also focused on a specific type of adaptive behaviour: suggesting items in a list that are likely to be selected by users as part of their ongoing task.

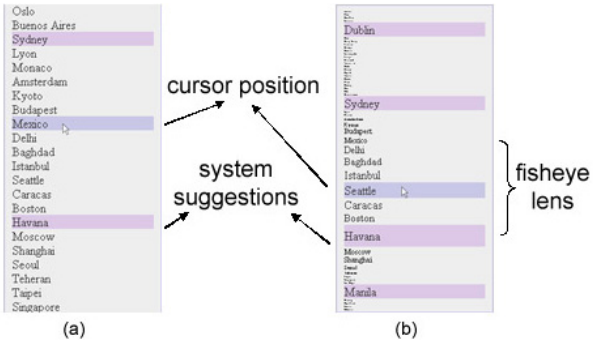

Figure 1. Tested techniques: (a) Highlighting suggestions (NORMAL), (b) Shrinking non-suggested items (SHRINK).

Our experiment tested two adaptation techniques, which were based on zooming and colour annotation. The two techniques are demonstrated in Figure 1. The first technique (NORMAL - Figure 1.a) simply highlights suggested items by changing the background colour. In addition to highlighting items, the second technique (SHRINK - Figure 1.b) shrinks non-suggested items.

The SHRINK technique is enhanced by a fisheye lens which allows users to explore minimized items, reducing the cost of incorrect system suggestions. Influenced by fisheye menus [1], the fisheye lens affects both the font size of the items as well as the height of their visualization. In our experiment, the fisheye lens contained 17

items in total. Their height ranged between 18 pixels, which was the normal height used by the high-context technique, and 3 pixels, which was the height of minimized items. The whole lists of items were visualized, i.e., no scrolling was used. Clearly, the SHRINK technique required significantly less space to visualize the same number of items. Depending on the number of suggested items, the height of a SHRINK list was 51-57% less than the height of a normal list. Adapting an information space by shrinking information that is non-relevant to a user's goals has been suggested by Tsandilas and schraefel [10] as a means of enhancing the focus on the user's task while preserving the surrounding context of information.

In our experiment, items were randomly sorted in the lists. This allowed us to examine adaptation in a general context in which either ordering does not obey a clear semantic relationship, or a particular sorting does not assist the task of the user. Adaptive behaviour is usually more valuable in such situations. The experiment examined adaptation techniques in separation from the inference mechanism used to decide on which items to highlight. The inference mechanism was considered as a black box which "somehow" highlighted items with respect to a given accuracy. As a result, participants could not predict the result of adaptation, as adaptation did not follow any clear pattern. We should, however, note that unpredictability is a common problem of most real adaptive interfaces [10].

Apparatus and Participants

The experiment was conducted in full-screen mode on a 2GHz P4 PC with 512 MB RAM running Windows XP. An 18-inch flat monitor was used at a resolution 1280 by 1024 pixels. Six female and six male volunteers, 24-40 years old, participated in the experiment.

Task

Participants had to complete a series of selection tasks. For each task, participants were presented on screen with the name of a country (*goal item*) and were asked to identify and select it from a list of country names. Each list contained 50 items randomly selected from a pool of 80 country names. Four or eight items were highlighted in the list by using either the NORMAL or the SHRINK technique. The goal item was either included or not included in the set of highlighted items with respect to the adaptation *accuracy* tested. We define accuracy as the percentage of tasks for which the goal item was highlighted. Immediately after the user selected the correct country name, a new task started and the experimental system requested the user to select a different country name. The set of highlighted items was constantly updated when a new task started. In the case of the SHRINK technique, a subtle animation effect was used to smooth the transition between such updates. Tasks were grouped into blocks of five tasks. For each block, in 5 out of 5 (100% accuracy), 4 out of 5 (80% accuracy), or 3 out of 5 (60% accuracy) tasks, participants were asked to select an item that had been suggested, i.e., highlighted, by the system.

Design and Procedure

A full factorial design with repeated measures was used. We varied three independent variables: (1) adaptation accuracy *Accuracy* (100%, 80%, and 60%), (2) adaptation technique *Technique* (NORMAL and SHRINK), and (3) number of system suggestions *Suggestions* (4 and 8 suggestions). The three variables were nested according to the above order. Each participant was exposed to all the 12 conditions, i.e., different combinations of the independent variables. A Latin square was used to balance the order in which participants tried the three accuracy levels. The order in which the instances of *Technique* and *Suggestions* were presented to participants was also balanced.

For each condition, participants completed 12 blocks of different selection tasks. After every four blocks, there was a brief pause, which allowed participants to relax. Participants had also to complete two practice blocks for each condition. All the participants completed the same blocks in the same order for all the 12 different conditions. However, following Findlater and McGrenere's approach [3], learning effects between different conditions were minimized by masking the adaptive list with a different set of country names for each condition. A pilot study indicated that long and well-known country names were usually located faster than short and less popular country names. In order to balance this effect among the 12 participants, we used 12 configurations of country names, which were assigned to the 12 conditions using a Latin square. This means that each configuration was applied to every condition exactly once, and no participant was exposed to the same configuration more than once. The selection of items that participants had to locate were randomly but almost uniformly distributed along the length of the list. The selection of suggested items was also randomly and uniformly distributed along the list.

The experiment was designed to fit in a single session, which was approximately 60-80 minutes long. At the beginning, participants were given a 3 minutes practice session. During this session, they were asked to locate a series of items in a non-adaptive version of both the traditional and fisheye list. The purpose of this practice session was to familiarize participants with the experimental setting and the fisheye lens used by the SHRINK technique.

With the exception of the 100% accuracy level, participants were not informed about the exact level of adaptation accuracy used in the experiment. The terms "low prediction accuracy", "high prediction accuracy", and "perfect prediction accuracy" were used instead. At the end of their session, participants were asked to give subjective estimates of the accuracy levels that they had experienced.

Measures

We examined three main dependent variables: (1) time *BlockTime* to complete a block, (2) time *SuggestedTime* to complete a task when the goal item had been highlighted by the system, and (3) time *NonSuggestedTime* to complete a task when the goal item had not been highlighted.

Since the number of successfully assisted tasks differed among accuracy levels, *SuggestedTime* was only measured for the subset of tasks in which adaptation was correct for all the experimental conditions. Likewise, *NonSuggestedTime* was only measured for the subset of tasks in which adaptation was incorrect for all the experimental conditions that corresponded to the 60% and 80% accuracy levels. In conclusion, there was one measurement per block for the first dependent variable, three measurements per block for the second dependent variable, and zero (100% accuracy) or one (60 and 80% accuracy) measurement per block for the third dependent variable. Errors made during the selection tasks were also recorded.

RESULTS

Data for a total of 8640 selection tasks were collected. As several participants reported, they frequently failed to identify the goal item within the set of suggested items. This situation caused large delays as participants realized their mistake after having searched the whole list. For this reason, we decided to isolate outliers generated for the *SuggestedTime* measure, replace them by the maximum non-outlier value encountered for the corresponding condition, and study them separately. We used 8 seconds as the cut-off value for such outliers since this resulted in a clear separation of the distribution of *SuggestedTime* into two distinct sets. 2.2% of the total number of measurements for *SuggestedTime* were identified as outliers.

Time Measures

In order to fix severe deviations of the distributions of the time measures from normality and ensure the reliability of our results, all the ANOVA tests were performed on the natural logarithm of the original measurements. The logarithmic transformation resulted in distributions close to normal. We denote by *lnXTime* the variable that results from *XTime* after applying the logarithmic transformation.

As shown in Figure 2, user performance degraded as the accuracy level became low. The main effect of *Accuracy* on *lnBlockTime* was found to be significant ($F_{2,22}$= *526.296, p<.0001*). Pair-wise comparisons showed significant differences for all the pairs of means (*p<.0001*). *Accuracy* had also a significant effect on *lnSuggestedTime* ($F_{2,22}$=*29.052, p<.0001*), which demonstrates that low accuracy levels affected users' trust on the system's suggestions. Pair-wise comparisons showed significant differences between the means (*p<.05*). We should note that the number of outliers (*SuggestedTime* > 8 sec) also increased as accuracy decreased. More specifically, we identified a total of 78 outliers for *Accuracy = 60%*, 36 outliers for *Accuracy = 80%*, and no outliers for *Accuracy = 100%*. On the other hand, the main effect of *Accuracy* on *lnNonSuggested* was not found to be significant ($F_{1,11}$=*1.669, p=.223, 21.9%* power).

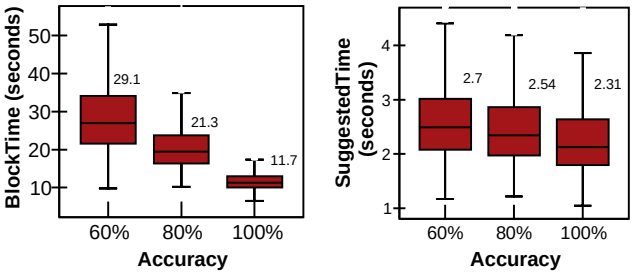

Figure 2. Boxplots demonstrating the effect of *Accuracy* on: (a) *BlockTime*, and (b) *SuggestedTime*. The numbers above the boxes show mean times.

Accuracy affected differently each of the two tested adaptation techniques. The effect of the interaction *Accuracy×Technique* on *lnBlockTime* was found to be significant ($F_{2,22}$=*20.890, p<.0001*). As Figure 3.a demonstrates, the SHRINK technique was slightly faster when accuracy was 100% (*p≤.002*, using Bonferroni's adjustment), since participants had to perform shorter mouse movements. Its performance, however, degraded faster than the performance of the NORMAL technique as accuracy decreased. This is due to the fact that SHIRNK was significantly slower than NORMAL in selecting items that were not suggested by the system ($F_{1,11}$=*103.734, p<.0001*).

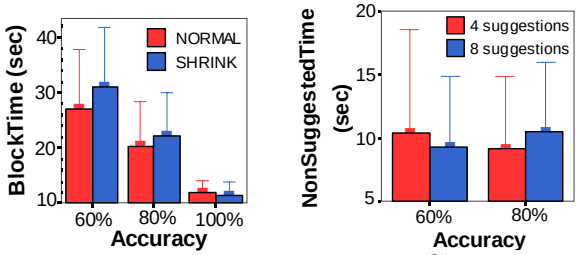

Figure 3. Graphs demonstrating the interactions (a) *Accuracy x Technique* on *BlockTime*, and (b) *Accuracy x Suggestions* on *NonSuggestedTime*. Standard deviations are shown.

As expected, the number of suggestions affected user performance. The mean time needed to select a correctly suggested item was significantly slower when 8 instead of 4 items were suggested ($F_{1,11}$=*211.09, p<.0001*). The number of suggestions did not have a significant effect on the mean time needed to select items that were not suggested by the system ($F_{1,11}$=*1.55, p=.238, 20.7%* power) . However, we observed an interaction effect between *Suggestions* and *Accuracy* ($F_{1,11}$=*12.959, p=.004*) as shown in Figure 3.b. According to some participants' comments, highlighted items often helped them to chunk the list and scan its items faster. Surprisingly, participants reacted

differently when the accuracy level was 80%. Apparently, they scanned suggested items more carefully in this case, and as a result, they spent more time when the number of suggestions was larger.

Errors

The SHRINK technique generated a large number of errors when the accuracy was imperfect. In more detail, a total of 105 errors were recorded in this case, as opposed to 42 errors recorded in the case of the NORMAL technique (60-80% accuracy). This can be explained by the fact that the motor space available for clicking on a non-suggested item was smaller when the SHRINK technique was used. Employing a lens-locking mechanism as the one used by fisheye menus [1] could address this problem.

Qualitative Results

Participants were asked to select among a range of accuracy levels that best estimated the proportion of times that the goal item was highlighted. They were asked to give separate estimates for the low-accuracy condition (60% accuracy) and the high-accuracy condition (80% accuracy). Although most participants gave a correct estimate of the high-accuracy level, the great majority of participants underestimated the level of low-accuracy. More specifically, 9 out of the 12 participants believed that more than 50% of the times the system failed to highlight the goal item. Several participants noted that they felt frustrated when the system's suggestions were regularly incorrect. It is likely that this fact radically decreased their confidence about the system's ability to correctly adapt the list. This result is consistent with previous research [7, 9], which suggests that user trust over automation reduces as automation becomes incompetent.

DISCUSSION AND CONCLUSIONS

We have presented a controlled experiment studying the performance of two adaptation techniques that suggest items in adaptive lists. The results indicate that the effectiveness of different adaptation techniques may vary depending on the accuracy of the prediction mechanism. Shrinking information that is likely to be irrelevant to a user's needs can be useful, since it reduces the size of the visualized space preserving on the same time valuable context information. This approach, however, was shown to delay the searching of items that had not been suggested by the system. As a result, the performance of the technique degraded as the accuracy of the system's suggestions became low.

An interesting result of our study is that accuracy affected not only the overall user performance but also the ability of participants to locate items that were correctly suggested by the system. This result can be explained by the decrease of user reliance on the system's suggestions as accuracy decreased and can be further justified by taking into account the participants' subjective responses.

The results do not directly show whether and when adaptive behaviour was effective in terms of user performance. Although answering this question was out of the scope of our experiment, we could argue that for the accuracy levels that we tested, adaptation was always effective. Even when accuracy was 60%, the mean time to locate an item using the slowest technique (SHRINK) was 6.3 seconds. Given that the mean time to locate a non-suggested item using the NORMAL technique was 8.73 seconds, and given the fact that we did not observe any main effect of the number of suggestions on the above measure, we can expect that the mean time to locate an item without adaptation to be present would not be faster. Experimentally validating thresholds under which adaptation techniques are useful is a goal worthy of future research.

ACKNOWLEDGMENTS

We thank Graeme Hirst and the anonymous CHI reviewers for their valuable comments. We also thank all the volunteers who participated in our experiment.

REFERENCES

1. Bederson, B.B. (2000) Fisheye menus. *ACM UIST*. San Diego, USA. p. 217-225.
2. Brusilovsky, P. (1996). Methods and Techniques of Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*. **6**(2-3). p. 87-129.
3. Findlater, L. and J. McGrenere. (2004) A comparison of static, adaptive, and adaptable menus. *ACM CHI*. Vienna, Austria. p. 89-96.
4. Fischer, G. (2001). User Modeling in Human-Computer Interaction. *User Modeling and User-Adapted Interaction*. **11**(1-2). p. 65-86.
5. Greenberg, S. and I. Witten. (1985). Adaptive personalized interfaces: A question of viability. *Behaviour and Information Technology*. **4**(1). p. 31-45.
6. Höök, K. (1997) Evaluating the Utility and Usability of an Adaptive Hypermedia System. *ACM IUI*. Orlando, USA. p. 179-186.
7. Lee, J.D. and K.A. See. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*. **46**(1). p. 50-80.
8. Mitchell, J. and B. Shneiderman. (1989). Dynamic versus static menus: An exploratory comparison. *SIGCHI Bulletin*. **20**(4). p. 33-37.
9. Tiernan, S.L., E. Cutrell, M. Czerwinski, and H. Hoffman. (2001) Effective Notification Systems Depend on User Trust. *INTERACT*. Tokyo. p. 684-685.
10. Tsandilas, T. and m.c. schraefel. (2004). Usable Adaptive Hypermedia Systems. *New Review of Hypermedia and Multimedia*. **10**(1). p. 5-29.