

Empirical Evaluation of Adaptive User Modeling in a Medical Information Retrieval Application

Eugene Santos Jr., Hien Nguyen, Qunhua Zhao, and Erik Pukinskis

Computer Science and Engineering Department
University of Connecticut
191 Auditorium Road, U-155, Storrs, CT 06269-3155
{eugene,hien,qzhao,erik}@cse.uconn.edu

Abstract. A comprehensive methodology for evaluating a user model presents challenges not only in choosing metrics, but also in assessing overall usefulness from both user and system perspectives. In this paper, we describe such a methodology and use it to assess the effectiveness of an adaptive user model embedded in a medical information retrieval application. With the methodology, we demonstrated that the user model helps improve the retrieval quality without degrading the system's performance and identified problems overlooked in the user model's architecture with regards to usability. The empirical data helps us to analyze weaknesses in our user model and develop potential solutions.

1 Introduction

Empirical evaluation of a user model's effectiveness is useful for both the user modeling community and the researchers in the area of the target application. Without a comprehensive methodology, the usefulness of an embedded user model can not be properly analyzed. Worse yet, researchers can be misled by the idea that adding a user model will improve system performance and reduce the users' workload when it actually degrades the system's performance resulting in important information needed by the users to be omitted. Unfortunately, empirical evaluation often has been overlooked even in the user modeling community itself. As pointed out in [2], only a third of the papers from the *User Modelling and User-Adapted Interaction* journal (1990 - 1999) included any type of evaluation. In particular, in information retrieval (IR), the usability of the applications enhanced by user models has been overlooked even though standard metrics have been well-established for retrieval performance [3]. While valuable in many respects, these metrics are not enough for assessing systems in which the primary goal is not to only maximize the quality of the retrieval process, but also maximize the quality of the user experience. Recently, the evaluation methods have shifted towards more concerns for the end users, including friendliness and responsiveness. Most of the existing methods which included usability evaluation focused solely on the appraisal of the interactions between users and graphical user interfaces (GUI) while ignoring the interactions between users and the information retrieved [1, 5]. In summary, the existing evaluation methods for

IR applications which use user models, focused either just on the accuracy of the retrieval process or just on the user interaction with a GUI. As such, they do not reflect the overall effectiveness of the user model and do not take into account the entire user experience with the system.

Here, we describe our methodology for evaluating the effectiveness of the user model in a medical IR application called Kavanah. The goal of the user model used in Kavanah is to accurately capture user intent and adapt the retrieval process by modifying the search query correspondingly [7].

We assess the effectiveness of the user model with regard to the target application in terms of its influence on system response time, accuracy and usability. We argue that by doing so, we obtain a more complete picture of the effectiveness of the user model from both user and system perspectives. The system response time assessment addresses our concern of whether the user model may degrade system performance. The accuracy assessment uses two common metrics, precision and recall [8], and ensures the user model improve retrieval quality while offering the benefits of easy comparison against other IR applications. Finally, our usability assessment combines both qualitative and quantitative metrics to evaluate the user’s experiences with a GUI and the information retrieved.

2 User model architecture in Kavanah

The goal of the user model is to accurately capture and represent a user’s intent. We partition user’s intent into three formative components: Interests, Preferences and Context. The Interests component captures the focus of the user’s attention. The Preferences component describes the actions that can be used to achieve the goals and is stored in a preference network (Figure 1c). The Context provides insight into the user’s knowledge behind the goals upon which the user is focused. It is stored in the user context network (Figure 1b).

Kavanah first accepts a natural language query from the user and converts it to a query graph (Figure 1a). The user model modifies the query graph based on the user’s interests, preferences and context having been captured (Figure 1d). The modified query graph is matched against each document graph and returns documents where the similarity is greater than a user-defined threshold. Note that each document is represented as a document graph which contains concept nodes (e.g Cancer) and relation nodes (e.g Isa). The user relevance feedback is used to update three components of the user model accordingly. For a more detailed information, please see [7].

3 Methodology and experimental execution

Kavanah is implemented as a web application. We started with a small database of 521 records to evaluate our methodology before scaling to a larger one.

Effects of the user model on system response time. Our hypothesis is that the user model will not significantly degrade system performance. We take 15 seconds as an acceptable delay time [9]. We test this hypothesis by assessing

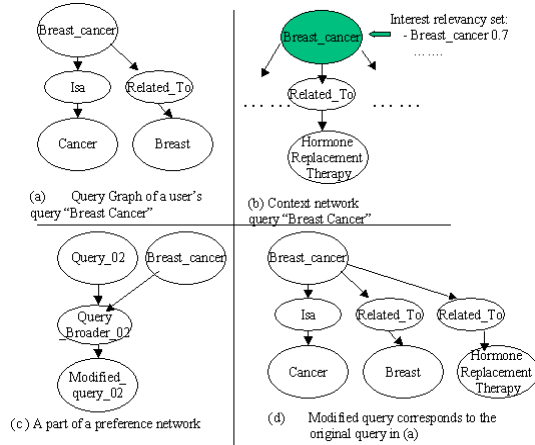


Fig. 1. Query graph, context, preference networks and modified query graph

the response time of Kavanah with and without the user model. The user model is constructed on-the-fly based on the relevance feedbacks collected during this test. The response time is defined as follows: $responsetime = T_{end} - T_{begin}$ in which T_{end} and T_{begin} are the time at displaying the results and the time at issuing the query, respectively. We ran this test with 2 sets, each set contains 50 natural language queries on some randomly chosen medical topics. We denote the values of response time for Kavanah running with a user model as $a_1..a_n$ and those without as $b_1..b_n$ ($n=50$). We then compute the slowdown rate for each test set as $slowdownrate = \frac{\sum (b_i - a_i)}{n}$. In this experiment, the average slowdown rate of these two test sets is 0.65 which means that by adding the user model, the system will slow down by 0.65 seconds. Note that the average response time for Kavanah without a user model is 40.65 seconds. This experiment demonstrated that adding a user model does not impact performance significantly.

Effects of the user model on system accuracy. Our hypothesis is that the user model in Kavanah will improve the accuracy of the retrieval process. We ran the system over a set of 35 queries with and without the user model. This set of queries is built around the topics such as cancer, diabetes, and infection on which our database has a lot of information. We set the threshold to filter out irrelevant documents to be low to maximize the retrieval of relevant documents available in the database. In our experiments, the precision and recall of the retrieval algorithm with a user model is at least as good as without a user model and improves as the user model learns from the user interactions. For example, by query 34 "Breast cancer risks for the elderly", without a user model, the retrieved results contained a lot of documents on breast cancer in general. With a user model, the system eliminated these documents and returned those documents focusing on breast cancer risks for women with hormone replacement therapy which reflected the user's goals in the query set. An important observation here is that the normal tradeoff of improved precision with decreased recall

did not occur in our case. With our user model, we are able to increase the precision, though small, while keeping the recall the same or even increasing. This trend opens a window of opportunity to fill in the gap to improve both precision and recall simultaneously through user modeling.

Usability evaluation. Our hypothesis is that the user model will not add extra workload to the users and will eventually reduce it. To measure the usability of Kavanah, we have adapted the framework laid out in [6]. We asked 16 undergraduate pre-med and pharmacy students to perform an extensive search on two topics, taking notes and providing feedback about relevant documents as they went. Our measure of usability has two components: objective task efficiency, or the ratio of task completion to task time, and subjective user efficiency, the ratio of user effectiveness to user effort. Task completion was calculated from the number of relevant documents the user identified and the number of topics they covered in their notes. Subjective user effectiveness and user effort were captured by a simple effectiveness questionnaire and the NASA Task Load Index [4], respectively.

After carrying out this experiment, we were unable to find a significant effect of the user model on efficiency, $t(15)=-.560$, $p=.585$, or task efficiency, $t(15)=-.047$, $p=.963$. There are several factors that may have contributed to this. Because of the small size of the database, participants were not motivated to issue many queries, which meant that Kavanah didn't have enough feedback to react to the user's behavior before they finished their task. Second, Kavanah's behavior is largely opaque to the user. It is our hypothesis that making the query expansion more interactive and more visible to the user will allow Kavanah to realize the usability potential that the AUI paradigm provides. Further research in this area is necessary.

4 Discussion and Summary

The unified framework of our evaluation methodology enabled us to jointly analyze the effects of the user model on the system and users, which may not have been possible if each evaluation is conducted in isolation.

In Figure 2, we show some possible outcomes of experiments and provide possible analyses of the effects of the user model on the system and user behaviors. We assume that the outcome of each experiment can be good (G) or bad (B) and denote effects on system response time, accuracy and usability tests as RT,A,U respectively. Outcomes of task efficiency and user efficiency are denoted as TE and UE respectively. Based on this table, we can determine if the user model has an overall bad or good effect on the system and users. Even better, we can use the real numbers obtained from these tests to analyze further about the actual causes of any effects and make decisions for possible solutions for the next design iteration. This analysis gives us a better understanding rather than looking at the outcomes of each test in isolation because it incorporates more information from different perspectives. Due to space limitations, we focus on

cases where the effects on system response time is good. The possible analyses are similar for the other outcome.

RT	A	U	Possible Analysis
G	B	TE: B, UE: B	The user model degraded system accuracy. Thus, the users spent a lot of effort with little success.
G	B	TE: G, UE: B	The user model degraded system accuracy but the user who knows the domain well managed to complete the tasks at the expense of extra effort.
G	B	TE: B, UE: G	The user model degraded system accuracy but the user did not know the domain as well as he thought.
G	B	TE: G, UE: G	The user model degraded system accuracy but the user who has a different domain view from the experts involved in system development still could find right information fast enough because he knows the domain well.
G	G	TE: B, UE: B	The user model did not help the expert user who has different domain knowledge from the experts involved in system development. Thus, he was frustrated for not finding any right information despite his knowledge. If the user is a novice, he might not know enough to complete the task in the given allotted time.
G	G	TE: G, UE: B	The user model helped the user who knows the domain well to retrieve the right information at the expense of extra effort because he and the experts involved in system development have different domain views.
G	G	TE: B, UE: G	The user model did not help the user, who thought he found the right information. In fact, he may not know the domain well enough to finish the tasks.
G	G	TE: G, UE: G	The user model supports users well.

Fig. 2. Possible analyses of influence on response time, accuracy and usability.

In summary, we have not only demonstrated that the evaluations actually address our concerns about whether the user model would delay the system or decrease the accuracy of the retrieval process, but we also pointed out how these empirical measures actually helped us identify the potential problems with the current architecture of the user model and with the target system.

References

1. Brajnik G.; Mizzaro S.; and Tasso C. 1996. Evaluating User Interfaces to Information Retrieval Systems. In *Proceedings of SIGIR 96*.128-136.
2. Chin N. David. 2001. Empirical Evaluation of User Models and User-Adapted Systems. In *User Modeling and User-Adapted Interaction*. Vol 11. 181-194.
3. Harter S. P.; and Hert C. A. 1997. Evaluation of Information Retrieval Systems: Approaches, Issues and Methods. In *Annual Review of Information Science and Technology*. Vol 32. 3-94.
4. Hart, S. G.; and Staveland, L. E. 1988. Development of the NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Hancock, P. and Meshkati, N. (Ed.). North Holland B.V., Amsterdam. 139-183
5. Koenemann, J.; and Belkin, N. 1996. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of CHI 96*.205-212.
6. McLeod, M.; Bowden, R., Bevan, N. 1997. The MUSiC Performance Measurement Method. *Behaviour and Information Technology*. Volume 16(4). 279-293.
7. Santos E. Jr.; Nguyen H.; and Brown M.S. 2001. Kavanah: An active user interface Information Retrieval Application. In *Proceedings of 2nd Asia-Pacific Conference on Intelligent Agent Technology*.412-423.
8. Salton, G.; and McGill, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.
9. Shneiderman, B. 1998. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley.