Open User Profiles for Adaptive News Systems: Help or Harm?

Jae-wook Ahn Peter Brusilovsky Jonathan Grady Daqing He Sue Yeon Syn
School of Information Sciences, University of Pittsburgh
135 N. Bellefield Ave.
Pittsburgh, PA 15256, USA
+1 412 6249404
{jaa38,peterb,jpg14,dah44,sus16}@pitt.edu

ABSTRACT
Over the last five years, a range of projects have focused on progressively more elaborated techniques for adaptive news delivery. However, the adaptation process in these systems has become more complicated and thus less transparent to the users. In this paper, we concentrate on the application of open user models in adding transparency and controllability to adaptive news systems. We present a personalized news system, YourNews, which allows users to view and edit their interest profiles, and report a user study on the system. Our results confirm that users prefer transparency and control in their systems, and generate more trust to such systems. However, similar to previous studies, our study demonstrate that this ability to edit user profiles may also harm the system's performance and has to be used with caution.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *Indexing method;* H.3.3 [**Information Search and Retrieval**]: *Information filtering; Relevance feedback;* H.3.5 [**Online Information Services**]: *Web-based services;* H.5.2 [**User Interfaces**]: *Graphical user interfaces (GUI).*

General Terms
Experimentation, human factors, performance.

Keywords
News personalization, user profile, open user model, control, trust

1. INTRODUCTION
Adaptive News Systems belong to one of the most popular type of personalized Web-based systems [3]. They are designed to help users access their daily portion of news over the Web. Taking into account user interests and preferences, adaptive news systems attempt to recommend the most interesting and relevant news items for individual users. One evaluation demonstrated the impressive effectiveness of adaptive news systems [4] and encouraged more work in this area. Over the last five years a range of projects focused on adaptive news delivery reported progressively more elaborated techniques for both user modeling and adaptation [1; 6; 7; 8; 13; 17].

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

However, along with more elaborated personalization techniques offering better performance, the adaptation process in these systems has become more complicated and thus less transparent to the users. As we discovered in the process of user studies in the context of the DARPA-supported GALE project (http://www.darpa.mil/ipto/Programs/gale/index.htm), it is hard even for professional information analysts to understand elaborated personalization mechanisms. As a result, the users felt frustrated and in less control of the situation and the system, and could not develop adequate trust to the personalized suggestions generated by the adaptive system.

Reading news is a specific type of information access, and we believe that information access is ultimately a human-controlled interactive process [12]. Through working with the systems, human users build up mental models about what they want, about the characteristics of the data collection, and about the functionalities of the systems. Through these models, the users guide their moves, anticipate results from the systems, and develop tactics and strategies to control the system and thus control the interactive access process. If the users fail to build up those mental models, they would feel frustrated and give less trust to the systems.

We believe that transparency can be applied to address the above problems of trust and control in adaptive news systems. In this paper, we concentrate on the application of open user models in providing transparency to adaptive systems. An adaptive system with an open user model shows the content of the user model to the user, so that the adaptive system becomes more transparent to the user. Moreover, a subcategory of open user models known as "editable user models" even allow users to change the content of the models to provide missing information or delete errors in the models. This effectively provides a mechanism for the user model to be examined and edited, and thus to "tune" the adaptation process. As a result, the user would feel more in control of system performance.

Open user models are relatively popular in the field of adaptive educational systems and their use in this area has demonstrated multiple benefits that these models can bring [5; 14]. However, models used in the field of adaptive news and, more broadly, personalized information access are different from user models in educational systems. User models for adaptive information access typically track user interests (in contrast to user knowledge in adaptive educational systems) and have relatively distinct structure. To stress the special nature of user models in adaptive information access systems these models are typically referred to as user profiles [10]. Although open and editable user profiles have not been discussed often in information retrieval (IR), the

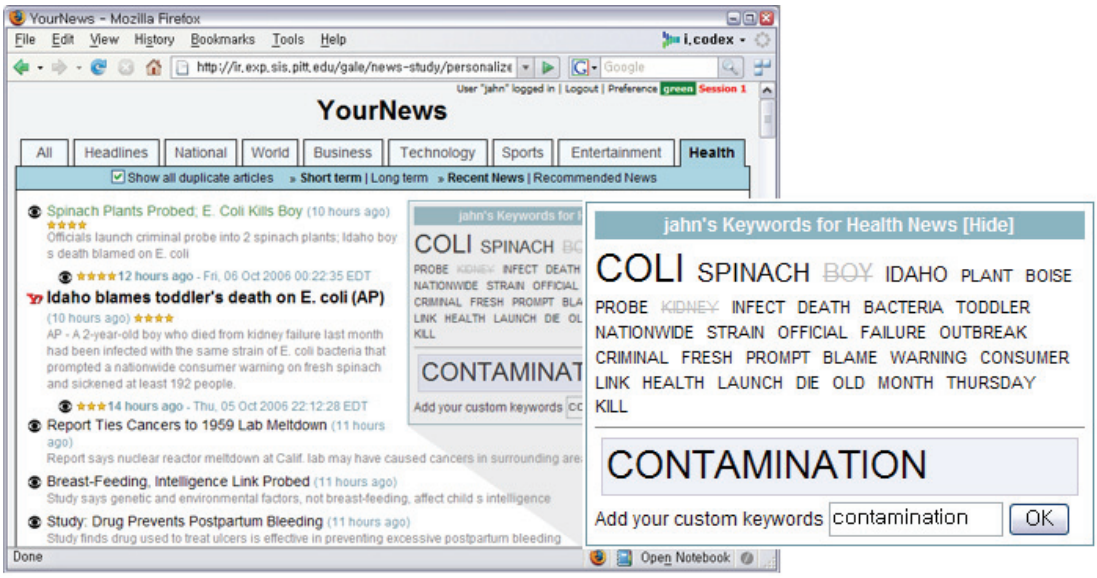

Figure 1 YourNews interface with the user model editor

idea of utilizing users' feedback to improve the performance of information systems is not new in IR. Relevance feedback (RF) has long been identified as an effective method for enhancing the performance of retrieval systems [20]. Ideas similar to open and editable user profiles have also been explored in the context of relevance feedback and query expansion. Studies conducted by Koenemann and Belkin [15] on different levels of the user's control on the expansion terms show that the increasing openness of the expansion terms and the higher level of the user's control on the query expansion improve search effectiveness. This includes the findings that 1) participants performed 15% better when they were able to view and manipulate the terms, 2) they used less iterations to develop equally good or better queries, and 3) participants had higher subjective views of the interface with more control. White and Marchionini tested a slightly different technique for allowing users to control the relevant feedback information [22]. In their setting, users were able to select a list of suggested additional query terms while the query is formulated. Their results show that the so-called real-time query expansion technique is more useful for exploratory search tasks – especially early in a search task – when users' needs may be most uncertain. However, they also acknowledge that the technique can lead the users down incorrect search paths, because users do not have enough information to know the effect of their actions.

There are very few examples of open and editable user profiles for information access [2] and almost no reported studies of open profiles. The only reported study of open user profiles for adaptive news access [21] brought rather interesting results. The study demonstrated that the ability to view and change their profiles is a mixed blessing in adaptive news context: it may help, but it may also harm. More specifically, user involvement can improve incomplete or bad profiles, but typically harms good ones. The study, however, has not provided a definite answer about the use of editable profiles because it explored a rather old-fashioned off-line profile editing. With off-line editing, the user does not see the results of the profile change immediately; instead, they are getting the news adapted to the edited profile with the next portion of news after some sizeable delay.

Despite the results from [21], we believe that transparency and control provided by open and editable user profiles are still valuable features, however, their usefulness needs the support of immediate feedbacks in adaptive information access. Because of lacking of fully understanding of the system, the collections, and sometimes the tasks, when presented the ability to modify the system's profiles, users may inevitably take some actions that would hurt the performance of the system. However, as long as the users can receive immediate feedback to their actions, they may learn from their mistakes, and hopefully make a better decision in the future. Therefore, although the performance of users' individual actions varies, their overall performance should be better, and their subjective satisfactions would be higher. Our previous studies on interactions in Cross-Language Information Retrieval demonstrated this [11].

In our work we attempted to explore the role of open and editable user profiles in a more appropriate context of on-line profile editing where the users can see the work of the new profile right after changes are made. We expected that the ability to get immediate feedback from the system would allow the users to change the profile appropriately and eliminate possible problems of this approach. This paper presents the results of our work. The next section introduces an adaptive news system, YourNews. The remaining parts of the paper report our study of editable user profiles performed with YourNews.

2. YourNews: A PERSONALIZED NEWS SYSTEM

2.1 Personalized News Presentation

YourNews (http://ir.exp.sis.pitt.edu/gale/news) is a Web-based system for personalized news access. Like many other adaptive news access systems [3], YourNews observes a user's news-reading behavior, constructs a user model (profile) representing user interests, and uses this model to recommend the most relevant news articles. YourNews assembles its content from 62 RSS news feeds from 8 sources. The collected information is organized into 8 topics (8 feeds per topic, on average) and presented to the user (Figure 1). Topics are separated (National,

World, Business, etc.) to avoid mixing together the user's interests in different areas. The system maintains a separate interest profile for each topic, and each user profile and set of news articles for a given topic are shown in separate tabs.

YourNews' crawlers periodically gather new articles from RSS feeds, passing them to an indexing module to build an index based on title, description, and content. The indexing module creates and stores weighted (TF-IDF) term vectors of the articles by the well-known vector space model [19]. Like other standard term-based recommendation systems, we extract tokens by dividing the text into terms by white spaces and special characters, remove stopwords such as articles and prepositions, and stem each term to be stored in the term vectors. In order to expose the user models (more specifically, the *terms* stored in the user models) to users, we adopted the Krovetz stemmer [16] that extracts term stems in more human readable forms, unlike other stemmers such as the Porter stemmer [18].

The user interest profile for each topic is also represented as a weighted prototype term vector extracted from the user's news view history. We collect N articles from users' past views, and the 100 top-weighted terms are extracted to generate the final prototype vectors. The recommendation process occurs by comparing these user models or prototype term vectors to new incoming article term vectors. Similarity scores between the user model and news articles are calculated using cosine coefficients ranging from 0 to 1, so that the target articles can be ordered by their similarity to the user model. The system maintains two kinds of interest profiles according to the time period the user model considers: *short* and *long-term*. *Short-term* profiles consider only the 20 most recently viewed news item, whereas *long-term* profiles consider all past views. Thus, each profile can express specific and general user interests, respectively. Given that news is separated into eight topics, 16 interest profiles exist in a single user model.

The user models are applied to generate two kinds of personalized views (or lists) for each topic: *recent news* and *recommended news*. The *recent news* view presents the 100 most recent news articles ordered by time. The *recommended news* view presents articles no older than one month ordered by relevance to the user topic profiles. For each article, the system presents the title and the subtitle extracted from RSS feeds. The title serves as a link to the full content of the article. Following the link opens the news article in a separate window and adds it to the list of viewed articles for user modeling purposes. In both the *recent* and *recommended news* views, the links to the articles are augmented with visual cues (Figure 1) - including star icons (1 to 5 stars), font weights, and font sizes (bold from 2 stars and larger from 3 stars) – to indicate the strength of the recommendations. The personalized view also applies a degradation function along with the conventional similarity computation. This process was adopted to stress more recent news articles, where recentness plays a greater role in the article's relevancy compared to other domains. For this task, a slightly modified sigmoid function is applied to the similarity scores calculated in the previous step, so that the articles older than one week from the user's access time should have half of the original similarity scores.

Users can select 9 topic tabs (8 topics plus 1 "All topics" tab), two different time periods (*long* and *short-term*), and two types of views (*recent* and *recommended news*). Therefore, they are provided with 36 views according to this combination. The default view is All (topics), *recent short-term news* (time period) view. When a user first starts using this system, it cannot provide any recommendations, because the user model is empty. Thus, YourNews initially behaves like a non-adaptive news aggregator, showing a list of news stories ordered by time. However, when the user follows a link and reads just one story, the system activates the recommendation process. It extracts terms from this viewed story and constructs profiles to be compared to candidate stories for recommendation according to the algorithm described above. If some stories displayed in the *recent news* mode have higher similarity than the threshold, relevant visual cues are added to the simple news list. When the user switches to the *recommended news* view, a list of articles sorted by the similarities to the user profile is displayed. The same rules are applied to YourNews' eight other topic tabs. The user can switch back and forth between two different user models, *long* and *short-term*. In this study, however, the *long-term* model did not play as great a role as it does when used conventionally, because the duration of the experiment was relatively short. Therefore, subjects did not have an opportunity to accumulate a long history of news views.

2.2 Open User Model and Transparency

Despite some technical innovations, the basic personalization architecture described in the previous section mostly follows the rather standard news recommendation approach used in several earlier systems [3]. The innovative part of YourNews is the open user model that was developed to increase the transparency of system personalization mechanisms and to provide some control over the personalization. The right-hand side of each topic view (Figure 1) shows the user model communication interface.

The interface *opens* the typically hidden profile of user interests by showing users the list of keywords that form their individual profiles. The font size of the keywords represents their relative importance or weight. For example, on Figure 1 the weights of the terms "COLI" and "SPINACH" in the profile are the highest. Using font size to stress the importance of an information fragment is a standard technology in the field of adaptive hypermedia. A similar approach is used on many social bookmarking sites for a different purpose – to show relative popularity of user tags.

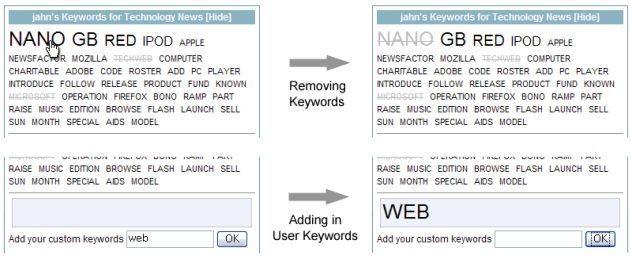

Figure 2 Editable user model: adding/removing keywords

The user model in YourNews is not only open, but also *editable*. Each term can be removed from the profile or re-enabled with a single mouse click. The removed keywords are shown in the profile in strike-through (for example "NANO" in Figure 2) allowing users to re-enable them later, but are not used for personalization purposes. Users can also add their own keywords (i.e. "WEB" in Figure 2), and can remove them later by clicking on the terms.

The personalization in YourNews is fully dynamic. Each interface event that may change the user profile (e.g., accessing a news

item or manual editing of the model) causes the news lists to be updated on the fly with all visual recommendation cues. Thus, users can immediately examine the effect of the changes, which, we expected, should lead to the improvement of the whole recommendation process.

Another new feature of YourNews that contributes to the transparency of the adaptation process and increases user control is the visualization of news items profiles. In order 1) to understand how a selection of a news story may affect the model, and 2) to predict which news stories may be affected by adding or removing terms from the open user model, the users should also know which terms are important and to what degree for each article. YourNews provides a way to show users the key terms contained in each article (Figure 3). When a user places the cursor over a story title, a popup window appears showing the important terms from the article. Terms are ordered by weight, and the same visual cues (i.e. font weight and size) used in the open user model viewer are applied here. Therefore, users can check the terms in each article, as well as their own terms in their user models, and can complete their tasks to find relevant news articles.

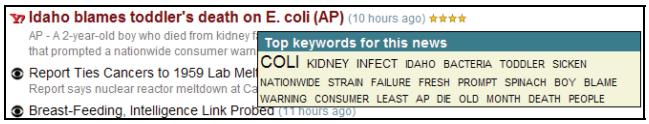

Figure 3 Displaying key terms for each article

3. THE STUDY DESIGN

In order to assess the value of YourNews' open user modeling features, an experimental study was performed using a full-fledged version of YourNews (i.e. with the open user model) as the *experimental system*. To create the *baseline system*, a second version of YourNews was created with full adaptive recommendation functionality (as presented in section 2.1) but with the open user modeling features presented in section 2.2 hidden from users.

We attempted to confirm two groups of hypotheses in the study:

H1: The experimental system, where users help the adaptation process by manipulating the open user model, performs better.

At an operational level,

H1-1: The experimental system will generate results with higher precision in terms of ordering

H1-2: Users of the experimental system will demonstrate higher task performance

H2: On the subjective level, users prefer the transparency of the user model in the experimental system

H2-1: Users are more satisfied with the experimental system

H2-2: Users actively change their profile

H2-3: Users appreciate the ability to view and change the profile

H2-4: Users appreciate the ability to view keywords behind news items

The designed scenario for the experiment was to have participants assume the roles of information analysts. Participants were asked to analyze news articles related to two specific topics, and collect articles reporting recent important events related to each topic. Participants were instructed to find relevant information pertaining to these important events, but were not specifically asked to find novel information. The definition of relevance in the experiment was explicitly mentioned to participants as "the degree of the importance of the content of a piece of text to be included in a report for a given topic".

Two topics were carefully chosen from events that occurred between September 25th, 2006 and October 15th, 2006. Documents for this time period were collected and the collection was frozen for the duration of the study. To allow the development of two independent profiles for the topics of interest, one topic was selected for the National tab and one for the Health tab. For the National topic, we selected "School Security" because several high-profile school shootings occurred throughout the U.S. during the time range. Users were asked to provide a status report including the relevant details of the shootings such as number of victims, where they occurred, public reaction, etc. For Health topic, we selected "Food Contamination" to cover events related to food safety issues in the U.S., particularly reports of E. Coli outbreaks linked to produce. Users were asked to provide a status report of possible outbreaks including the possible sources of the outbreak, related deaths or illness, etc.

To simulate "tracking" of information related to the topics of interest, the search tasks were split into two sessions, simulating two points of access to the data collection separated by 10 days. The first session provided access to news as participants would have seen on October 6th, 2006, and the second session provided access as of October 15th, 2006. The actual experimental sessions were held on separate days but within the same week (October 16th to 23rd.) From the task point of view, two sessions represented a more realistic scenario where in the second session the users were expected to identify information not available during the first session. From the user modeling point of view, the first session with a smaller number of relevant documents available served as a training stage (for both the users and their models), while the second allowed for the assessment of user performance over the whole set of documents.

During each session participants were expected to collect links to news items that they may need to include into a report to a superior, along with a representative passage from each item. To make the passage collection seamless, participants used the Google Notebook extension for Firefox Web Browser (http://www.google.com/notebook) that was preconfigured for the study. At the end of the second session, participants were able to edit and rank the list of collected passages. The edited list represented the result of their work, and was evaluated for both precision and recall.

Ten graduate students from the School of Information Sciences, University of Pittsburgh volunteered for the experiment. They were familiar with the information search task, but none was specifically interested in either of the study topics before the study. Each participant worked with both topics. Participants were randomly assigned a system to use for a given topic, and performed their search tasks on that system in both sessions. This design allowed for some control of between-subject differences and for a direct comparison of the two systems.

The study procedure is shown in Table 1. The study design was based on a within-subjects design with two groups of users performing tasks on two different systems. Before the first session,

participants were given a brief description of the system and tasks. Each session consisted of two search tasks (one on each topic) with a brief post-questionnaire and break between tasks. Participants were given 15 minutes to extract relevant information, and additional time was provided to rank extracted passages in their Google Notebooks. Including time for instruction, breaks, and filling out questionnaires, each session lasted approximately one hour.

Table 1 Experiment procedure

Session 1		Session 2	
User Group A	**User Group B**	**User Group A**	**User Group B**
Instruction and Training		Instruction and Training	
Health (Baseline)	Health (Experimental)	National (Experimental)	National (Baseline)
Questionnaire		Questionnaire	
Break		Break	
National Experimental	National Baseline	Health Baseline	Health Experimental
Questionnaire		Questionnaire	

4. RESULT ANALYSIS AND DISCUSSION

4.1 The Ground Truth

To measure system and user performance, we adopted precision (the fraction of selected items that are really relevant) as a major measuring tool. We established the "ground truth" for each topic by manually annotating the 456 news articles presented to at least one user during the experiment. Each article was rated on a 3-point scale (0=irrelevant, 1=marginally relevant, 2=fully relevant). Three annotators were involved in this task and cross validations on 90 samples were performed beforehand to improve the reliability of the annotation process.

With the manual annotation information by human annotators (ground truth), the overall system performance was calculated as in Table 2. Here, "Relevant" means the corresponding article matches well with the topic; "Marginally relevant" means linguistic matches occur (i.e. appearance of key terms), but the content is not related to the topic; and "Not relevant" means no relation of any nature to the topic. All items recommended by the system during the experiment were checked with the ground truth and classified into one of these three categories. For the National topic, a total of 919 recommendations (one item can be recommended several times in different situations) were made to the subjects, 712 of which were relevant (78%), 53 marginally relevant (6%), and 154 not relevant (17%). We can observe considerable topic dependence as Health topic shows lower performance than National topic. This might reflect the domain characteristics of the Health topic (e.g. terminology), where subjects tend to be unfamiliar and lack the knowledge to determine what information or keywords are relevant.

Table 2 Overall system performance

Topic	Relevance	Count	Ratio
National	Not relevant	154	0.168
	Marginally relevant	53	0.058
	Relevant	712	0.775
Health	Not relevant	253	0.398
	Marginally relevant	36	0.057
	Relevant	346	0.545

4.2 System Performance Analysis

The job of a personalized news system is to push the most relevant items to the top of the recommended news list. Having relevant items at the very top of the list (top 10 or top 20) is especially important since Web users are known to pay most attention to the first screen of results. To measure the performance of the adaptive recommendation in YourNews we calculated *precision at rank 10 and 20*. Top 10 and 20 news stories were collected from the "recommended news list" mode and checked into which category in Table 2 they fell: Relevant, Marginally relevant, or Not relevant. Therefore, precision at rank 20 means the number of relevant items from the rank 1 to 20 divided by 20. Thus, rank 20 represents approximately one screen where users can see those articles without scrolling down to the next page.

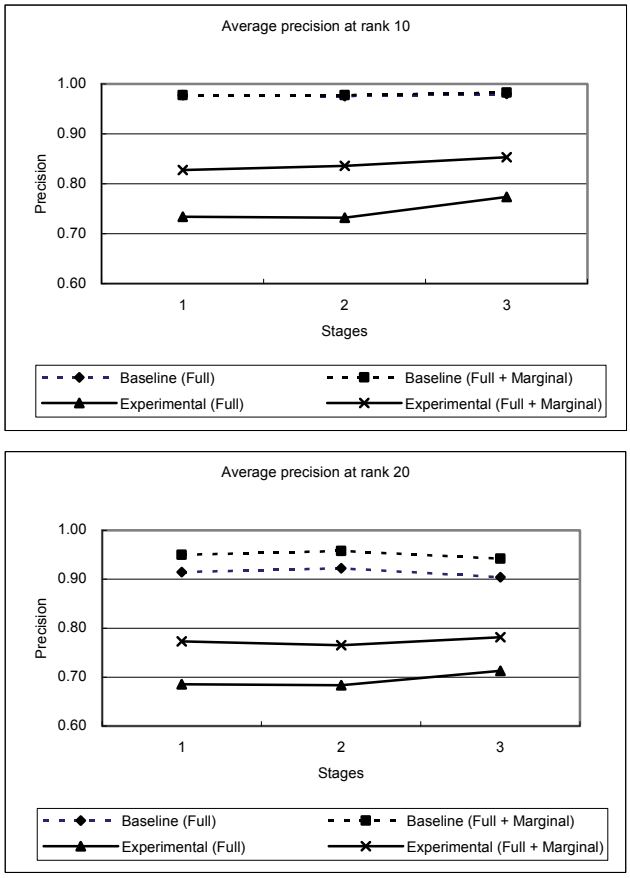

Figure 4 Average precision at rank 10 and 20

Figure 4 shows the temporal change of the system performance during session 2. We divided the entire session into 3 stages and measured the precision of each document recommendation screen encountered by the user during the stage at rank 10 and 20 (as noted above, new ordering is produced after every article access and every profile change). These 3 stages were defined by dividing the total number of tasks the subjects made by 3. For example, if a subject was shown with 9 recommendation lists in a single session, stage 1 ranges from the 1st to the 3rd list, stage 2 from the 4th to the 6th, and stage 3 is from the 7th to the last. In these recommendation lists, users are able to quickly find the most relevant articles to their needs by looking at highly ranked items. This precision was averaged for each user and then among

all users. The results are presented in Figure 4. Here "Full" means the precisions were calculated using only the fully relevant items from the manual annotations and "Full + Marginal" means those were achieved using fully and marginally relevant items, which is a less conservative approach.

From these graphs, we can see that the baseline system performance is better than the experimental system, on average, and the difference between them is statistically significant (independent t-test, p=0.000). The baseline system without the open user model shows precisions above 0.9 for this session. Even the least conservative measures (precision at rank 10, Full+Marginal) did not exceed the worst performance of the experimental system. Even though there were some improvements with the experimental system at stage 3 they were very minimal.

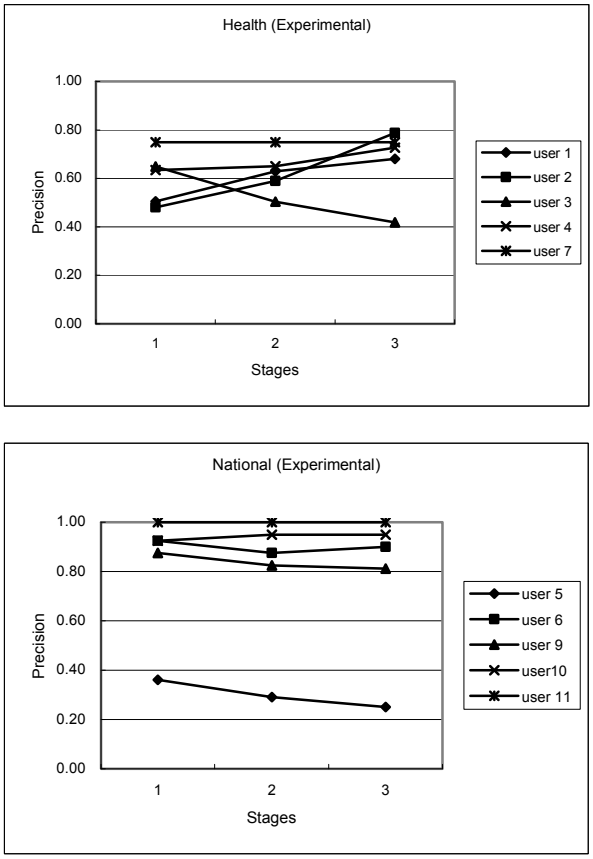

Figure 5 Per-subject system performance for two topics

This initial result contradicts our first hypothesis; that is, if we provide users with open user models and allow them to add or remove keywords from the user models, then system performance will increase – users might know well what the good and bad keywords are in order for them to achieve their goals. By looking deeper at the system performance data extracted during the study, we found out that the users and topics were not identical to each other. We can see this phenomenon in Figure 5, which represents the experimental system's performance during 3 stages of the second session (per subject). We separated the users by the topics they worked on: Health and National. What we observe is that the overall system performance is different according to the topics. Health shows much lower performance than National topic. The performance of the former topic ranges from 0.4 to 0.8 whereas that of the latter shows far better performance (> 0.8), with the exception of one case, "user 5". This subject started from a far worse score (< 0.4), the lowest between topics and systems, and then dropped below 0.3.

For the Health topic, "user 3" displayed a similar behavioral pattern. In this topic, the best initial performance by a user was around 0.8. Users who had performed lower initially tended to improve their performance toward this level as the stages progressed. However, "user 3" started from a relatively lower precision at around 0.6 and the performance continued to decline to 0.4, the lowest score in the corresponding group.

In order to find out how the performance of these two subjects negatively affected the system's performance, we can think about three perspectives: user's personal characteristic, topic (Health or National) and system type (baseline or experimental). Even though it appears system performance is dependent on topic – the experimental system performed better with the National topic than with the Health topic – the subjects with distinctly lower system performances had similar behavior in both topics. Also, the system performances of the baseline systems with these users were relatively good (0.95 and 0.79 respectively), so the explanation that the subjects were not good at working with the adaptive news system would be incorrect. The last factor that could affect system performance with them is the system type, more particularly the open user model in the experimental system.

Table 3 Keyword manipulation with the open user model vs. performance

Topic	Subject	Add	Remove	Total	P@20
Health	User 1	4	5	9	0.610
	User 2	6	0	6	0.607
	User 3	1	4	5	0.516
	User 4	6	0	6	0.673
	User 7	3	0	3	0.750
National	User 5	1	15	16	0.304
	User 6	2	1	3	0.900
	User 9	3	4	7	0.838
	User 10	0	0	0	0.943
	User 11	0	0	0	1.000

Since the ability to add and remove keywords was the main difference between systems, we examined the keyword manipulation frequency with the open user model and compared it with the system performance (Precision at rank 20) individually (Table 3). A quick glance at the data provides a hint about the connection between the amount of user model changes and the resulting system performance. We notice that "user 5", with whom the system performance recorded the lowest, performed many keyword manipulations with the open user model. The user removed 15 keywords and added one. Similar to before, the user's average precision is only 0.304, which is less than one-third of the best average precision 1.0. The next subject in question, "user 3" did not perform as many manipulations as "user 5", but still removed the second highest number of keywords in the group. "User 1" did more keyword additions and removals than "user 3", and the system performance ranked 3rd out of 5 subjects in the group. The highest performances in each group ("user 7" and "user 11") are 0.750 and 1.000 respectively. Neither subject removed any keywords, and "user 7" added only a relatively small number of keywords (3).

We found that the keyword manipulation frequency is correlated to the system performance; that is, the actions may have harmed

the performance. In order to examine the correct relationships we used a scatter plot and linear regression (Figure 6). The horizontal axis is the sum of adding/removing keyword frequencies of the subjects and the vertical axis measures the system performance, similar to Table 3. Squares and the dashed line represent the relationships between the frequency and the performance with the National topic, and triangles and the solid line are for the Health topic. We observe that negative relationships exist for both topics – more manipulation, poorer performance – and the effect was larger for the National topic, as evident by its greater slope. We checked the fitness of the plots to the lines with R-square, which denotes their degree of convergence to the lines. The values were 0.16 and 0.94 for Health and National, respectively, and the results show that we can have more confidence on the fitness with the data from the National topic.

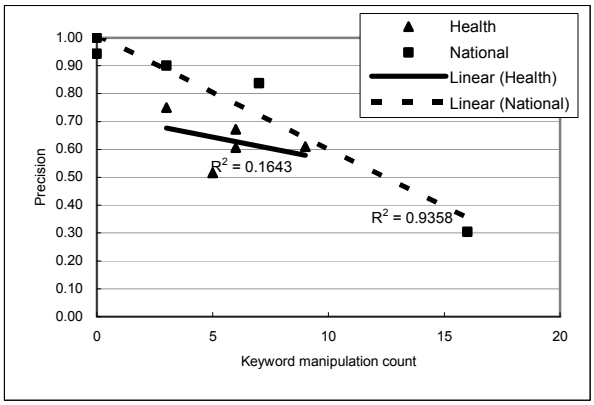

Figure 6 Keyword manipulation versus system performance

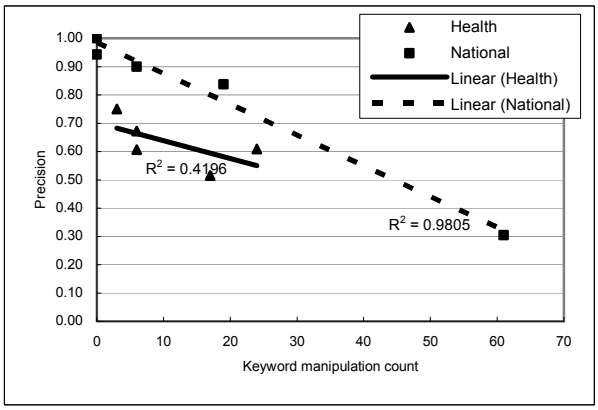

Figure 7 Examining keyword manipulation type magnitude: with 4 times more weights to removing than adding

In order to check which action (addition or removal) affected in what degree to these negative relationships, we tried to increase the weights of keyword removal frequency for calculating the sum of the frequencies (*add_frequency + weight * remove_frequency*) used in the graph. As we increased the weight, the R-square values increased proportionally, but this tendency went down when the weight reached more than 4 (0.42 vs. 0.16 and 0.98 vs. 0.94 with the removal frequency weight = 4) (Figure 7). It provides some evidence that keyword removals harmed the performance about 4 times more than adding keywords.

4.3 User Performance Analysis

As described in the previous section, we asked the subjects to collect links and passages for their topic reports. We assume that better system support in recommending relevant documents will result in better quality reports. The quality of these reports was measured by standard precision and recall that were adapted to the context. The precision of the user report selections was calculated against the ground truth, as number of relevant items (we showed fully relevant items only) divided by number of selected items (see Table 4). First of all, it shows a similar tendency with the system performance discussed before. The performance of the experimental system is lower than the baseline system and the National topic performance is slightly higher than the Health topic.

Table 4 User annotation results (Precision)

Topic	Precision
Health	0.80
National	0.83

System	Precision
Baseline	0.87
Experimental	0.75

Table 5 User annotation results (Recall)

Topic	System	Average Recall
Health	Baseline	0.750
	Experimental	0.600
National	Baseline	0.778
	Experimental	0.711

The recall values were calculated taking into account that, for the task assigned to the user, collecting information about all major events related to the topic was important. To calculate the recall, human annotators divided two main topics into subtopics representing main events (about 9 for each topic). The items selected by each subject were categorized into these subtopics, and then the topic-level recall was calculated as a fraction of topics covered in the report. As shown in Table 5, the results were again consistent with the main trend: the topic performance was better for National topic and the group performance better for the baseline system.

4.4 User Action Analysis

Until now, we have examined how the system performance changed according to user behaviors when they were working with the baseline and the experimental system with the open user model. Examining the log data allowed us to see what the users were actually doing. Table 6 shows how many documents they viewed per each session.

What we were interested to check is how well the system worked for the user; that is, whether users trusted the recommendations provided by the system. Because the system performance and the user's actions are closely related to each other with adaptive systems, it is not easy to clearly determine the exact extent of the trust of the users and the misinformation of the system. However, we can check some aspects to examine these variables: time spent reading articles (TSR), and average ranks of user clicks. The average rank of clicks provides some evidence of user's personal

impressions or the system ordering mechanisms and of user trust in the system. Low rank or user selection (corresponding to accessing items high in the recommended list) tells us that from the user's point of view the system ordering was good and they trusted this ordering. Time spent reading can provide some further evidence. High TSR indicated that the items selected by the user were considered relevant for more careful reading. Low TSR indicated that the selected item was not as relevant as expected and was, at the most, skimmed.

Table 6 Number of document views by users

System	Session	Number of views
Baseline	1	147
	2	138
Experimental	1	136
	2	131

We traced how much time users spent reading (TSR) each news articles. Users can select any document, but they can either keep reading it thoroughly or close it right away if they notice it is not what they wanted. Therefore, we can assume the misinformation (recommended but not useful) of each click by examining TSR. Table 7 shows the ratio of low TSR clicks, which are less than 5 or 10 seconds. In the baseline system, subjects spent less than 10 seconds with about half of the whole click. In the experimental system they spent more time with the articles they opened in session 1 (lower ratio of low TSR) but it suddenly increased in session 2 up to the maximum value among the low TSR statistics. That is, the subjects were able to invest more time reading the articles in the session 1, but they discovered more articles that were not useful in session 2. This corresponds to the low system performance of the experimental system in session 2, where the system precision was very low. Improper ordering might lead to the misinformation to users causing them to try many irrelevant items.

Table 7 Ratio of TSR clicks (in seconds)

System	Session	TSR<5	TSR<10	TSR≥10
Baseline	1	26.4%	49.3%	50.7%
	2	23.0%	48.9%	51.1%
Experi-	1	12.4%	33.3%	66.7%
mental	2	32.8%	54.6%	45.4%

Table 8 Average rank of clicks

System	Session	Stage	Average Rank
Baseline	1	1	28.6
		2	32.9
		3	28.6
	2	1	9.0
		2	13.4
		3	21.6
Experimental	1	1	13.9
		2	19.4
		3	27.0
	2	1	17.5
		2	13.1
		3	20.8

Another statistics we can use to the issue of trust and misinformation is the average rank of items (in recommended news lists) clicked by subjects. In Table 8, the data is presented per stage as we examined the system precisions. Overall, the ranks are rather low, especially lower than rank 20, which is the number of articles users can see in one screen. Even though higher ranks appear in the first stage of each session, soon the ranks decrease below rank 20, which means the subjects scrolled down the first screen and opened the corresponding article. Because items are arranged by their recommendation scores, the low rank means the subjects did not fully follow the system's recommendations and this provides some evidence for the misinformation and trust issue. The data also hint that the users' trust in system ordering in session 2 was about the same for both systems despite the experimental system producing relatively poor ordering and frequently causing the user to try documents that were not useful.

4.5 User Feedback Analysis

Following each search task, subjects were given a post-questionnaire to assess their satisfaction with the system (Table 9). For all questions, subjects were asked to rate their level of agreement from 1 (Extremely) to 5 (Not at all). For both systems, subjects were asked to rank topic familiarity, sufficiency of news, trust of system, control of system, and overall satisfaction. For only the experimental system, subjects were asked to rate the utility of the user model controls for adding, removing, and displaying terms.

Table 9 Post-questionnaire questions
(* indicates experimental system only.)

1.) Were you familiar with this topic before the search?
2.) Did the system provide you with sufficient news for your task?
3.) Were you confident in the system's ability to find useful information on this topic?
4.) Did you feel you had enough control over how the system recommended news items?
*5.) Did you find that **adding** terms was useful in helping the system find useful news items for this topic?
*6.) Did you find that **removing** terms was useful in helping the system find useful news items for this topic?
*7.) Does **displaying** the terms of your interest model help you understand how the system finds useful news items for this topic?
8.) Overall, did you have a positive experience with this system?

Chi-square tests were performed on the questionnaire data to determine significant differences in user answers by system and by topic. Table 10 and Table 11 show the mean responses for each session by topic and system, respectively, with overall means reported. As shown in Table 10, subjects indicated they were better able to find sufficient news for the National topic versus the Health topic in Session 1 ($\chi^2(3) = 9.086$, $p = 0.028$), but not overall. Although subjects felt they were more familiar with the national topic ($\mu = 3.1$) than the health topic ($\mu = 2.2$), which might help support these findings, the difference was not significant.

As shown by Table 11, the users' feedback regarding the experimental system was generally more positive than the baseline system's feedback. Moreover, for the summative

questions 2-4, the user attitude to the experimental system increased in session 2 while the attitude to the baseline system decreased. However, none of the differences were significant.

Table 10 Mean post-questionnaire responses to Questions 2-8, summarized by session and topic. (* p <= 0.05)

Q#	Session	National	Health
2	1	*4.30	*3.40
	2	3.80	3.70
	Overall	4.05	3.55
3	1	4.10	3.30
	2	3.90	3.70
	Overall	4.00	3.50
4	1	3.00	3.30
	2	3.20	2.80
	Overall	3.10	3.05
5	1	4.20	4.00
	2	3.00	2.80
	Overall	3.60	3.40
6	1	3.00	3.00
	2	3.60	2.60
	Overall	3.30	2.80
7	1	3.60	3.80
	2	3.40	3.60
	Overall	3.50	3.70
8	1	3.70	3.70
	2	3.60	3.60
	Overall	3.65	3.65

Table 11 Mean post-questionnaire responses to Questions 2-8, summarized by session and system. (* p <= 0.05)

Q#	Session	Experimental	Baseline
2	1	4.00	3.70
	2	4.10	3.40
	Overall	4.05	3.55
3	1	3.50	3.90
	2	4.00	3.60
	Overall	3.75	3.75
4	1	3.10	3.20
	2	3.40	2.60
	Overall	3.25	2.90
5	1	*4.10	N/A
	2	*2.90	N/A
	Overall	3.50	N/A
6	1	3.00	N/A
	2	3.10	N/A
	Overall	3.05	N/A
7	1	3.70	N/A
	2	3.50	N/A
	Overall	3.60	N/A
8	1	4.00	3.40
	2	3.80	3.40
	Overall	3.90	3.40

For the between-sessions comparison, the only significant finding was the decreased satisfaction from session 1 to 2 with adding terms to the user model ($\chi^2(3) = 8.333$, $p = 0.04$) assessed by Question 5. For Question 6 (removing terms), although there is no significant difference between session 1 and 2 feedback, overall subjects were rather neutral on the ability to remove terms from their profile. Finally, subjects moderately appreciated the ability to see the underlying profile (Question 7), and indicated in exit interviews that seeing the keywords behind each news item was more useful than the ability to see their entire profile.

The data suggest that, over time, subjects had less appreciation for the abilities to view and change their profiles in the experimental system. In exit interviews, many subjects also said that although they felt as though they had more control with the experimental system than the baseline, manipulating the terms directly did not give them enough control. They expected the addition and removal of terms to have a greater impact on the system's recommendations at the end of the first session and throughout the second session, when they wanted to find articles containing new information. Instead, many subjects were frustrated with the system presenting them redundant articles at the top of the recommended news view, and thus selected articles not recommended by the system.

5. DISCUSSION AND FUTURE WORK

Despite our expectations, our study didn't confirm that the ability to view and edit user profiles of interest in a personalized news system is beneficial to the user. On the contrary, it demonstrated that this ability has to be used with caution. Our data demonstrated that all objective performance parameters are lower on average for the experimental system. It includes system recommendation performance as well as precision and recall of information collected in the user reports. Moreover, we found a negative correlation between the system performance for an individual user and the amount of user model changes done by this user. While the performance data vary between users and topics, the general trend is clear – the more changes are done, the larger harm is done to the system recommendation performance. There is also some evidence that removing terms from the profile may harm the performance more than adding terms.

However, despite performance problems, the users preferred the experimental system at average in several aspects and rated their positive experience with this system higher than the experience with the baseline system. The average rank of user selections in the experimental system was relatively high in both sessions, which may indicate some reasonable level of trust in system performance. A combination of less perfect system performance with good level of trust may harm the effectiveness of user work with the system as indicated by lower precision and recall, as well as a relatively large number of short visits to news items.

The visibility of the user and items profiles was positively evaluated, however the user attitude to the ability to remove terms was rather neutral and their enthusiasm about the ability to add terms shown in the first session dropped significantly in the second session which is, actually, the only visible sign of user discouragement with the system.

The results of our study confirmed the controversial results of Waern's study [21]: the ability to change established user profiles typically harms system and user performance. We expected that YourNews' ability to support interactive profile changes would alter the situation, giving the user a chance to learn how to modify their profiles appropriately by immediately observing the results

of these changes, but it has not happened. Most likely, this ability has not provided a difference because the users were simply unable to distinguish good and bad system performance, just as they were not able to distinguish good and bad profiles in Waern's study. Users' relatively high ratings and trust in the experimental system's recommendations provide some evidence in favor of this hypothesis.

It is important to stress that like the majority of systems in its class, YourNews was focused on relevance of recommendation and uses simple techniques to control the novelty of recommended items. Our observations show that it was a serious problem in the task performed by our subjects. While our system was able to recognize duplicated news items to certain degree, a good percentage of the top recommended news were repeated information that are already known to subjects, which make them useless. In this context, some user model changes may be caused by user attempts to push already known items back, harming the system performance. We think that our current data still leaves the possibility that the ability to change the user model will be beneficial in a system that does a better job promoting novel items. We are currently working on the novelty control in YourNews and plan to run further studies of the improved system. This is another challenging problem with only one known attempt to handle it in the context of personalized news systems [9].

6. ACKNOWLEDGMENTS

This paper is partially supported by DARPA GALE project and by the National Science Foundation under Grant No. 0447083.

7. REFERENCES

[1] Ardissono, L., Console, L., and Torre, I. An adaptive system for the personalised access to news. AI Communications, 14 (2001), 129-147.

[2] Baudisch, P. and Brueckner, L. TV Scout: Lowering the entry barrier to personalized TV program recommendation. In: P. De Bra, P. Brusilovsky and R. Conejo (eds.) Proceedings of 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2002), Málaga, Spain, May 29-31, 2002, pp. 58-68.

[3] Billsus, D. and Pazzani, M. Adaptive news access. In: Brusilovsky, P., Kobsa, A. and Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Springer, Berlin, 2007.

[4] Billsus, D. and Pazzani, M.J. User modeling for adaptive news access. User Modeling and User Adapted Interaction, 10, 2-3 (2000), 147-180.

[5] Bull, S. Supporting learning with open learner models. In: Proc. of 4th Hellenic Conference on Information and Communication Technologies in Education, (Athens, Greece, September 29 - October 3, 2004), 47-61.

[6] Chen, C., Chen, M., and Sun, Y. PVA: A self-adaptive Personal View Agent. Journal of Intelligent Information Systems, 18, 2-3 (2002), 173-194.

[7] Conlan, O., O'Keeffe, I., and Tallon, S. Combining adaptive hypermedia techniques and ontology reasoning to produce dynamic personalized news services. In: Wade, V., Ashman, H. and Smyth, B. (eds.) Proc. of 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2006), (Dublin, Ireland, June, 2006), Springer , 81-90.

[8] Díaz, A. and Gervás, P. Personalisation in news delivery systems: Item summarization and multi-tier item selection using relevance feedback. Web Intelligence and Agent Systems, 3, 3 (2005), 135-154.

[9] Gabrilovich, E., Dumais, S., and Horvitz, E. (2004) Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: Proceedings of 13th international conference on World Wide Web (WWW'2004) New York, USA, ACM, pp. 482-490

[10] Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. User profiles for personalized information access. In: Brusilovsky, P., Kobsa, A. and Neidl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Springer, Berlin, 2007, in press.

[11] He, D., Oard, D.W., and Plettenberg, L. Studying the Use of Interactive Multilingual Information Retrieval. In: Proc. of Workshop on New Directions of Multilingual Information Access at SIGIR 2006.

[12] He, D., Oard, D.W., Wang, J., et. al., A. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi. ACM Transactions of Asian Language Information Processing, 2, 3 (2003), 219-244.

[13] Jokela, S., Turnpeinen, M., Kurki, T., Savia, E., and Sulonen, R. The Role of Structured Content in a Personalised News Service. In: Proc. of 34th Hawaii International Conference on System Sciences, 2001, 1-10.

[14] Kay, J. Scrutable adaptation: Because we can and must. In: Wade, V., Ashman, H. and Smyth, B. (eds.) Proc. of 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2006), (Dublin, Ireland, June 21-23, 2006), Springer Verlag, 11-19.

[15] Koenemann, J. and Belkin, N.J., A case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. in *Proceedings of CHI '96*, 1996, 205-212.

[16] Krovetz, R. Viewing Morphology as an Inference Process. In: Proc. of ACM SIGIR'93, 191-203.

[17] Magnini, B. and Strapparava, C. User modeling for news Web sites with word sense based techniques. User Modeling and User Adapted Interaction, 14, 239-257 (2004).

[18] Porter, M.F. An algorithm for suffix stripping. Program, 14, 3 (1980), 130-137.

[19] Salton, G. (ed.) The SMART retrieval system: experiments in automatic document processing. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[20] Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41* (4). 288-297. 1971

[21] Waern, A. User involvement in automatic filtering - an experimental study. User Modeling and User Adapted Interaction, 14, 201-237 (2004).

[22] White, R.W. and Marchionini, G. Examining the Effectiveness of Real-Time Query Expansion. *Information Processing and Management*. 2006