Impacts of User Modeling on Personalization of Information Retrieval: An Evaluation with Human Intelligence Analysts

Eugene Santos Jr. Qunhua Zhao, Hien Nguyen, and Hua Wang

Computer Science and Engineering Department
University of Connecticut
191 Auditorium Road, U-155, Storrs, CT 06269-2155
{eugene,qzhao,hien,wanghua}@engr.uconn.edu

Abstract. User modeling is the key element in assisting intelligence analysts to meet the challenge of gathering relevant information from the massive amounts of available data. We have developed a dynamic user model to predict the analyst's intent and help the information retrieval application better serve the analyst's information needs. In order to justify the effectiveness of our user modeling approach, we have conducted a user evaluation study with actual end user, three working intelligence analysts, and compared our user model enhanced information retrieval system with a commercial off-the-shelf system, the Verity Query Language. We describe our experimental setup and the specific metrics essential to evaluate user modeling for information retrieval. The results show that our user modeling approach tracked individual's interests, adapted to their individual searching strategies, and helped retrieve more relevant documents than the Verity Query Language system.

1 Introduction

It is both critical and challenging for analysts to retrieve the right information quickly from the massive amounts of data. The task of designing a successful information retrieval (IR) system for intelligence analysts is especially difficult, considering that even when given the same search task, each analyst has different interests and almost always demonstrates a cognitive searching style that is different from others analysts. Clearly, a user model of a intelligence analyst is essential to assisting the analyst in his/her IR task. Since the early 80s, user modeling has been employed to help improve users' IR performance [3]. In our recent efforts, we developed a dynamic user model that captures an analyst's intent in order to better serve his/her information needs [14, 15] in an IR application.

In order to properly assess the effectiveness of a user model, we need to measure how an analyst's performance and experience with an IR system are affected. Intelligence analysts are personnel for collecting and compiling information for government, law enforcement and defense, etc. They are trained be self-conscious of their reasoning process [12], which includes the IR process. One major barrier for

evaluating a system designed for analysts is the limited accessibility to working intelligence analysts, and the nature of the information used in the evaluation.

To assess the effectiveness of our user modeling approach, we have conducted an evaluation with three working intelligence analysts. The objectives of this evaluation are: 1) to evaluate how our user model enhanced IR system performs when compared against a traditional IR system implemented with a keyword based query language, the Verity Query Language (VQL) [16]; 2) to study the impacts of our user model on augmenting personalization in IR; and, 3) to get feedback from the evaluators on user performance. The results show that our user modeling approach tracked the analyst's intents and adapted to the individual analyst's searching styles which helped them retrieve more relevant documents, especially those relevant to each analyst than the system implemented with VQL.

This paper is organized as follows: We first briefly present related work on user modeling in IR and its evaluation. Next, our user modeling approach is described, followed by our prior work on IR evaluation. Our evaluation methodology is then presented and our results are reported. Finally, we present our conclusions and future work.

2 Background and Related Work

The main objective of IR is the retrieval of relevant information for users. It is not an easy task, not only because of the explosive amounts of available information (especially unstructured information), but also the difficulty in judging the relevance, which can be objective or subjective in nature (reviewed by Borlund [2]). User modeling techniques attracted much attention in efforts at building a system for personalizing IR [3, 14]. However, proper evaluation of the user model for IR remains a challenge [4, 10, 17].

In the IR community, various methodologies, procedures, and data collections for evaluation of IR performance have been developed. In a typical experiment with data collections like Cranfield [5], a set of relevant documents is picked up by human assessors for a certain query (topic). Their judgments are considered to be objective [2]. The sets of relevant documents are then used for calculation of precision and recall. The criticism is that these experiments ignore many situational and mental variables that affect the judgment on relevance [8].

Besides applying metrics developed in the IR community, such as precision and recall, for measuring the effectiveness of the user model for IR [4, 7], efforts have also been made to study the impacts of the different systems on user behaviors. The emphasis was on the interaction between the user and the system. In a study done by Koenemann et al [7], the influence of four interfaces, which offered different levels of interaction in relevance feedback supported query formulation, to the user searching behaviors are studied. They found that different interfaces shaped how the users constructed their final queries over the course of the interaction. When the users could view suggestions and had control on the final actions, they needed less iterations to form good queries.

Recently, researchers in the user modeling community have focused on the development of general frameworks to conduct usability tests, which involves various forms of aptitude tests, cognitive tests and personality tests through surveys and questionnaires [4]. In IR domain, these results should be carefully considered, since previous research showed that user preference is not correlated with human performance [17]. Therefore, reliable conclusions could not be obtained solely based on either performance or user satisfaction. As Chin [4] pointed out, the difficulty lies with the evaluation approaches and study with real users to justify the overall effectiveness of a user model.

We attempted to evaluate our user modeling approach for IR, which is described in the next section, by measuring improvement in system performance, system adaptation to the user, and the user's experience with the system.

3 IPC User Model

Our user modeling module consists of three components: Interests, Preferences and Context, which is referred as the IPC model [14, 15]. Interests capture the focus and direction of the individual's attention; Preferences capture how the queries are modified and if the user is satisfied with the results; and Context provides insight into the user's knowledge. We capture user Interests, Preferences and Context in an Interest set, a Preference network and a Context network, accordingly. Interest set is a list of concepts, each of them associated with an interest level. Initially determined based on the current query, it is then updated based on the intersections of the retrieved relevant documents. The Preference network is captured in a Bayesian network [11], which consists of three kinds of nodes: pre-condition nodes, goal nodes and action nodes. Pre-condition nodes represent the environment in which the user is pursuing the goal. Goal nodes represent the tools that are used to modify a user's query; and action nodes represent how the user query should be modified. The Context network is a directed acylic graph that contains concept nodes and relation nodes. It is created dynamically by finding the intersections of retrieved relevant documents. The user model captures the analyst's intent and uses this information to modify analyst's query proactively for the IR application, please see [14, 15] for details.

The user model module has been integrated into an IR system. In the IR system, a graph representation for each document (called a document graph) is generated automatically in an offline process. The document graph is a directed acyclic graph consisting of concepts and the relations between concepts [14]. The query is also transformed into a query graph, which is then matched against each document graph in the collection. To speed up the matching process, only 500 documents that contain at least one term that exists in the query will move into the graph matching process. If there are more than 500 such documents, then the documents containing less terms from the query will be removed. The similarity measure between document graph and query graph is modified from Montes-y-Gòmez et al [9], also see [14].

4 Evaluation Methodology

Previously, we evaluated our user modeling approach by using evaluation measures, procedures and data collections that have been established in the IR community [10]. These experiments demonstrated that our user modeling approach did help improve the retrieval performance. It offers competitive performance compared against the best traditional IR approach, Ide dec hi [13], and offers the advantage of retrieving more quality documents quickly and earlier [10].

As such, we would like to compare our user model enhanced IR system to a more traditional system implemented with a keyword based query language. Furthermore, we would like to have an opportunity to study the impacts of our user modeling approach on augmenting personalization within the IR process, and get feedback from real intelligence analysts about their personal experience. A data collection from the Center for Nonproliferation Studies (CNS, Sept. 2003 distribution. http://cns.miis.edu/) has been chosen as the testbed for this evaluation. It contains 3,520 documents on topics of country profiles concerning weapon of mass destruction (WMD), arms control, and WMD terrorism. It was chosen because its content and its built-in commercial query system from Verity, Inc. [16] that can be used as a baseline system for comparison. In the following text, we will refer to our user model enhanced IR system as the UM system, and CNS with VQL as the VQL System.

The evaluation took place at a laboratory of the National Institute of Standards and Technology (NIST) in May, 2004. The UM system package, which includes the pre-processed CNS database, was delivered to and installed at the NIST laboratory. Three evaluators, who are naval reservists currently assigned to NIST with intelligence analysis background participated in the experiments. Since only three analysts were available, to obtain some fair comparison data, we have to run the UM system and the VQL system side by side during the evaluation. The same queries were input into both systems and the retrieved documents compared. For the VQL system, analysts needed to note on paper which documents were relevant to their interests for each query; for the UM system, in addition to recording the relevancy, they were asked to mark checkboxes beside the documents if they were relevant ones. There was a short tutorial session to show the analysts how to work with the UM system, such as indicating the relevancy. For the VQL system that has a graphic user interface (GUI) similar to Google, it is straightforward to use.

The experimental session lasted about 4 hours for each analyst due to analyst availability and laboratory scheduling. Participants were asked to carry out a searching task on “research and development in Qumar that supports biological warfare” (Note that some of the location names have been replaced). Because of this timing constraint, the participants were asked to check the first 10 returned documents for relevancy only, and the task was limited with just 10 fixed queries (Table 1). For any empirical study, one challenge lies in the large numbers of variables to control (including the human factors). By scripting the queries, we can avoid introducing more variables into our experiments, such as different queries, different number of query inputs, and error in natural language processing. It allowed us to have a better control on the experiment in such a short session, and focus on the main objectives of the evaluation, which is to study the impacts of user model on the IR, as described in the introduction. The queries were extracted and modified from a database that

collected other intelligence analysts' IR activities at the NIST laboratory, which allows us to construct a realistic evaluation session. The UM system started with an empty user model, which means that the user model initially knew nothing about the analyst, and had to start learning about the user from the very beginning.

Table 1. The 10 queries used in the evluation experiments

1	Qumar research biological warfare
2	Qumar research institute, university biological warfare
3	Qumar biological research and biological warfare
4	Biological research facilities in Qumar
5	Intelligence assessment on Qumar biological research
6	Qumar foreign connections in biological weapons program
7	Bacu, Qumar and Russia connections to WMD
8	Qumar's biologists visits Bacu
9	Russian biotechnology, missiles, aid to Qumar
10	China supply and Qumar biological weapons program

Besides the IR task, analysts were asked to fill out an entry questionnaire about their background and experience with searching programs; and, respond to an exit questionnaire about their experience on working with the UM system.

5 Results

The experience in intelligence analysis for the three participants ranged from 5 months to 7 years. Two of them use computers as a tool in their analysis job, while one does not (Table 2). They all felt comfortable with using search tools like Google, and considered themselves well-informed on the topics of WMD and terrorism. Analyst 3 stated that he has never used a system that requires feedback for annotating relevancy (Table 3). The most interesting observation is that the three analysts tend to take different approaches in IR. Analyst 2 looks at the big picture first; while analyst 3 likes to start with the details. Analyst 1 practices a mixed approach that depends on his knowledge of the topic. If much was already known, then he would try to create an outline of the useful information; otherwise, he would look for some details first (Table 3).

After 4 hours, two analysts finished 10 queries that we provided, and Analyst 3 finished 9 queries (Table 4). All of them managed to identify more relevant documents when working with the UM system than they did with the VQL system (Table 4). The precision were 0.257 and 0.312 for the VQL system and the UM system respectively. Since a document could be returned and identified multiple times as relevant for different queries, we also counted the numbers of unique (or distinct) documents that have been returned by the system and found as relevant by each participant. The data showed that when they were using the UM system, each of them was presented with more unique documents, and selected more unique documents as relevant (Table 4). The total number of unique relevant documents for all 10 queries

returned by the UM system is 39, while the number is 27 by the VQL system, a 44% increase (Table 5).

The number of documents selected as relevant by more than 2 analysts are 15 in the UM system and 19 in the VQL system, respectively. Notice that the number of documents marked as relevant by just one analyst is 24 when using the UM system, while this number is only 12 for the VQL system (Table 5). This suggests that more information that is specifically relevant to each analyst's individual interests had been retrieved by the UM system. By using the UM system, the analysts displayed their differences in identifying the documents that were relevant to their individualize interests and searching style.

Table 2. Demographic data

	1	2	3
Highest degree	JD	MS	BA
Length of time doing analysis	7 years	5 years	5 months
Computer expertise	novice	medium	medium
Use computer to do analysis	not at present	yes	yes
Experience doing queries	yes	yes	yes
Query expertise	novice	medium	medium

Table 3. Questions on information seeking behaviors of three participants.

	1	2	3
What is your overall experience with systems using ranked outputs and full-text databases, such as Google? (1-7) 1 is very experienced, 7 is no experience	3	1	1
Have you ever used a system that asked you to indicate whether a document or other system response was relevant? Yes, No	Y	Y	N
When faced with a search problem do you tend to: (a) Look at big picture first, (b) Look for details first, (c) Both	c	a	b
What is your knowledge of Terrorism (1-7) 1 very experienced, 7 no experience	2	3	2
What is your knowledge of WMD? (1-7) 1 very experienced, 7 no experience	3	2	2

By the end of the experiment, the analysts were asked to fill out the exit questionnaire. Generally, they agreed that the scenario used in the evaluation experiment was very realistic, and gave an above average score for feeling comfortable at preparing a report on their task after querying for information. When asked about the system performance and their satisfaction, they scored the UM system as above medium (3.7/5.0) (Table 6 and 7). Notice that they felt the UM system was somewhat demanding, especially in mental effort and the temporal effort. Since relevancy assessment is a mentally demanding process by itself, and the analysts were required to finish the experiment in about 4 hours, which included 10 queries (i.e., more than 100 documents to review, of which some of them may be quite long), and

working with 2 different systems at the same time, we think this is a result of the workload the analysts had in the experiments. As the data shows, the UM system presented more unique documents to the analysts, and helped analysts retrieve more relevant documents. In particular, it helped them retrieve more information that is relevant to their individual interests, which suggests that the user model was tracking the user's personalized interests.

Table 4. Number of documents presented to the analysts, and number of documents marked as relevant with each of the systems.

	VQL system			UM system		
Analyst	1	2	3	1	2	3
Documents presented	100	100	90	100	100	90
Relevant documents	11	31	33	16	41	36
Unique document presented	49	49	45	67	72	54
Unique relevant documents	9	19	21	10	29	23

Table 5. Unique relevant document retrieved by two systems.

	UM System	Verity System
Total unique relevant documents	39	27
Documents marked as relevant by all 3 analysts	8	3
Documents marked as relevant by more than 2 analysts	15	19
Documents marked as relevant by only 1 analyst	24	12

6 Discussion and Future Work

In this paper, we present our evaluation methodology and the results for our user modeling approach. Since the ultimate goal of IR is to meet the user's information needs, testing by actual end users (the analysts in this case) is an evaluation that can not be replaced by other methods. The involvement of end users can help us avoid problems with traditional IR evaluation metrics which excludes the user's individual information needs. Our evaluation answered the question on impacts of user modeling on the retrieval performance of an IR system by measuring the number of unique documents presented to the analysts and relevant ones have been identified; and studied the impacts of user modeling on the personalization of IR by tracking the difference between the documents retrieved by different analysts. By combining these results, we can judge if the user modeling is actually follows the user's individual interests, and improve the IR performance.

Intelligence analysts are trained experts specialized in IR and information analysis in certain areas. It is very hard to get time from real analysts to test a system in a experimental setting. We are very glad that we have had the chance to perform such an evaluation. Since the experimental time is limited (4 hours), we used a short scripted query sequence to reduce the number of variables in the experiment, which allows us to focus on our main objectives.

Table 6. Average score for performance of the UM system (1)

Question	Score
How realistic was the scenario? 1-5, 1 is not realistic, 5 is realistic	4.7
Did it resemble tasks you could imagine performing at work? 1-5, 1 not realistic, 5 realistic	3.7
How did the scenario compare in difficulty to tasks that you normally perform at work? 1-5, less difficult, 5 more difficult	2.7
How confident were you of your ability to use the system to accomplish the assigned task? 1-5, 1 less confident, 5 more confident	3.0
Given that you were performing this task outside of your standard work environment, without many of your standard resources, were you comfortable with the process of preparing your report? 1-5, 1 less comfortable, 5 more comfortable	3.7
Given that you were performing this task outside of your standard work environment, with access to a restricted set of documents, were you satisfied with the quality of the report/answers that you were able to find for this scenario? 1-5, 1 not satisfied, 5 satisfied	2.7

Table 7. Average scores for performance of the UM system (2)

Question	Score
How satisfied are you with the overall results for this task using OmniSeer? 1-7, 1 most satisfied, 7 least satisfied	4.3
How confident are you with the results that they cover all possible aspects of the task? 1-7, 1 most confident, 7 least confidence	4.7
Regarding this task, do you think the OmniSeer approach helped you to retrieve critical documents earlier in the process than the Verity system? 1-7, 1 strongly agree, 7 strongly disagree	3.7
Please rank the following factor: mental demand 1-7, 1 little 7 high	5.3
Please rank the following factor: physical demand 1-7, 1 little 7 high	2.0
Please rank the following factor: temporal demand 1-7, 1 little 7 high	5.0
Please rank the following factor: performance demand 1-7, 1 little 7 high	4.7
Please rank the following factor: frustration 1-7, 1 little 7 high	5.3
Please rank the following factor: effort 1-7, 1 little 7 high	6.0

Although there were only 3 analysts tested on the system within a limited period of time, the results are encouraging. First, the UM system provided more information to the analysts (returned more unique documents, which is usually can only achieved by asking more queries), and helped them to identify more relevant information. Second, even more importantly, experimental results suggest that the UM system tracked the individual interests of the different analysts, and returned different sets of documents to them individually. We know that the 3 analysts employ different seeking approaches (look for general information first, or look for details first, or use a mixed approach). With the UM system, the query was modified based on the user's

feedback. During the IR process, different analysts considered different documents as relevant based on their own knowledge, experience and goals, which led to the difference in the modification of the queries by the user modeling module. As a result, they were presented with different documents. This demonstrates the impacts of our user model on augmenting personalization in the IR process. With the VQL system, there is no effort to meet the individualized information needs. It is always the same set of documents returned for the same query. Because of the timing constraints, the evaluation only involved one task consisting of 10 queries. Also, the query sequence was fixed. If there were more queries asked freely by the participants, and with a larger database, the UM system would have been able to indicate even more significant differences among the analysts.

VQL is a very successful commercial query language. It has been developed and enhanced over more than a decade. Many advanced functions have been included in VQL, such like proximity, density, frequency, field, concept, word stemming, and word location [16]. It is obvious that our UM system, as a prototype, lacks many advanced features offered by the VQL system. For example, VQL's word location function helps the user find the keywords in the query (or words closely related) by highlighting them in the documents; the UM system does not provide the same convenience although we tried to implement a similar GUI with the intention of minimizing the interface differences. VQL uses keyword or concept indexing to accelerate searching process, which also has a big advantage over the current version of our UM system. These features could affect the evaluation outcomes, and might make the participants feel that the VQL system takes less effort.

In a study by Alpert et al [1], it has been pointed out that users want to feel that they are in control. In our case, the analysts were given a short training session and brief introduction on how to use the UM system before the experiment, and were informed that their feedback will be used by the system to try and improve performance. Unfortunately, it is far less than what is necessary. More work is needed in the future to help users understand how and why the system evolves and behaves, which will grant them more of a sense of being in command, and help users overcome suspicious attitudes, such as a system's ability to do it well enough to be useful.

Currently, in the UM system, relevance is explicitly selected by the analysts at the whole document level. When selected, the whole document is placed into the relevant set. This may introduce noise into the user model, since it is possible that only part of the document is considered relevant by the user. In the future, a system may be implemented with both explicit and implicit feedback mechanisms. Implicit feedback, like Hijikata's work [6], can both lessen the burden of marking the relevancy by the user and also identify the specific part that is of interest within the presented text. Explicit feedback can let the user be in control and indicate to the system what the most important relevant information is. We hope, with more concise feedback, our user model can better infer the user's intent and then assist their information needs.

Acknowledgments : This work was supported in part by the Advanced Research and Development Activity (ARDA) U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government. Also, we express our special thanks to Dr. Jean Scholtz and Emile Morse, who helped organize the

evaluation, collected data, and provide the preliminary data analyses. Without their help, this evaluation experiment would not have been successful. This work is a part of the Omni-Seer project, which involves Global InfoTek, Inc., University of Connecticut, University of South Carolina, and KRM, Inc. [15].

References

1. Alpert, S.R., Karat, J., Karat, C-M., Brodie, C., Vergo, J.G. : User Attitudes Regarding a User-Adaptive eCommerce Web Site. User Modeling and User-Adapted Interaction 13 (2003) 373-396
2. Borlund, P. : The Concept of Relevance in IR. Journal of the American Society for Information Science and Technology 54(10), (2003) 913-925
3. Brajnik, G., Guida, G., Tasso, C. : User Modeling in Intelligent Information Retrieval. Information Processing and Management 23(4) (1987) 305-320
4. Chin, D. : Empirical Evaluation of User Models and User-Adapted Systems. User Modeling User-Adapted Systems 11(1-2), (2001) 181-194.
5. Cleverdon C. : The Cranfield test of index language devices. (1967) Reprinted in Reading in Information Retrieval Eds. 1998. Pages 47-59.
6. Hijikata Y. : Implicit User Profiling for On Demand Relevance Feedback. 2004 International Conference on Intelligent User Interfaces (IUI04). ACM presses, Funchal, Madeira, Portugal. (2004) 198-205
7. Koenemann, J., Belkin, N.: A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In Proceedings of CHI 96 (1996) 206-212
8. Large, A., Tedd, L.A. Hartley, R.J. : Information Seeking in the Online Age: Principles and Practice. (1999) London: Bowker Saur
9. Montes-y-Gòmez, M., Gelbukh, A., Lpez-Lpez, A. : Comparison of Conceptual Graphs. In Proceedings of MICAI-2000, 1st Mexican International Conference on Artificial Intelligence. (2000) Acapulco, Mexico.
10. Nguyen, H., Santos, E. Jr., Zhao, Q., Lee, C. : Evaluation of Effects on Retrieval Performance for an Adaptive User Model. AH2004: Workshop Proceedings - Part I. Third Workshop on Empirical Evaluation of Adaptive Systems. (2004) 193-202
11. Pearl, J. : Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. (1988) Morgan Kaufmann, San Mateo, CA
12. Heuer, R. : Psychology of Intelligence Analysis. (1999) Government Printing Office.
13. Salton, G., Buckley, C. : Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science. 41(4), (1990) 288-297
14. Santos, E., Jr., Nguyen, H., Brown, S.M. : Kavanah: An Active User Interface information Retrieval Agent Technology. In Proceeding of the 2nd Asia-Pacific Conference on Intelligent Agent Technology. (2001) 412-423.
15. Santos, E., Jr., Nguyen, H., Zhao, Q., Wang, H. : User Modelling for Intent Prediction in Information Analysis. In Proceedings of the 47th Annual Meeting for the Human Factors and Ergonomics Society, Denver, Colorado, (2003) 1034-1038
16. Verity White Paper: The Verity K2 Discovery Tier, The Importance of Advanced, Effective Search Tools. (2004)
http://www.verity.com/pdf/white_ papers/MK0348c_WP_Discovery.pdf
17. Wilkinson, R., Wu, M. : Evaluation Experiments and Experience from Perspective of Interactive Information Retrieval. In Working notes of Empirical Evaluation of Adaptive Systems workshop at Adaptive Hypermedia Conference. (2004) 221-230.