CAPTURING USER INTENT FOR INFORMATION RETRIEVAL

Hien Nguyen, Eugene Santos Jr., Qunhua Zhao and Hua Wang
Department of Computer Science and Engineering
University of Connecticut
{hien,eugene,qzhao,wanghua}@engr.uconn.edu

Gathering information and making decisions based on retrieved information are the important tasks that every intelligence analyst has done daily. User modeling techniques have been exploited to help analysts to search for information effectively. In order to justify the effects of any user modeling techniques on helping analysts retrieve quality documents relevant to their tasks, we need to have a comprehensive evaluation method, which assess the improvement of retrieval performance and user performance. In this paper, we describe our evaluation of a cognitive user model for information retrieval with regards to retrieval performance. Our user model captures user intent dynamically by analyzing behavioral information from retrieved relevant documents for improving the retrieval performance and user's performance. In this evaluation, we assess the user model's short-term effects on a single query, and the user model's long-term effects on the whole search session. We compare our approach with the best traditional approach for relevance feedback in information retrieval, the Ide dec-hi, which is the approach of modifying queries using term frequency from relevant/non-relevant documents. We use the oldest collection of information retrieval on aerodynamics called CRANDFIELD. The results of this evaluation show that by exploring user intent, we achieve competitive performance in the feedback run compared to Ide dec-hi, at the same time our user model approach offers the advantages of retrieving more quality documents at the initial run compared to the term frequency inverted document frequency (TFIDF) approach. Our results have shown that our user modeling approach can be used to improve efficiency, learnability and interactivity of an information retrieval system.

Introduction

Storing and retrieving data from heterogeneous resources on the computers are the main tasks of every intelligence analyst's daily job description. With the increase of online and offline resources, these tasks have become mentally demanding and time consuming tasks. User modeling techniques have been exploited to help users, including analysts, improve their performance and improve the retrieval process since the late 80s (Brajnik et al 1987). A comprehensive evaluation framework to assess the effectiveness of a user model for IR is crucial to justify the effects of the user model on helping analysts achieve their goals. It should be able to answer if a user model improves retrieval performance and satisfies analysts simultaneously. In this paper, we present an evaluation of a cognitive user model for information retrieval (IR) with regards to retrieval performance. This is one important phase of an ongoing three-phase evaluation, which includes the evaluation of the accuracy, the effectiveness in terms of improving retrieval performance, and user performance of our user model (Nguyen 2004). With our user model, we can predict the goals and intentions of an analyst in order to better serve his searching processes by pro-actively retrieving novel and relevant information as it arises. In this evaluation, we compare our approach with the best traditional approach for relevance feedback in IR, the Ide dec-hi (Salton and Buckley 1990) using CRANFIELD collection (Cleverdon 1967). The Ide dec-hi is a classical IR approach that modifies a user's query by adding/subtracting the weights of terms from relevant documents/nonrelevant documents. CRANFIELD is the oldest collection for the IR community which contains 1398 documents on aerodynamics. Even though the evaluation is done with a hypothetical user, our results have some important implications from human factor perspective. We demonstrate that the efficiency, learnability and interactivity between a user and an IR system can be potentially improved using our user modeling approach. Our evaluation method provides the ability to compare with existing IR techniques while offering the ability to assess the special features of our model.

This paper is organized as follows: We begin by reviewing important related work in IR and User Modeling (UM) communities regarding the evaluation of a user model. Then we briefly overview our user model approach. Next, the experimental setup is presented and followed by the analysis of the results obtained. Finally, we present our conclusions and future work.

Background and Related Work

Our user model has made use of relevance feedback and query expansion techniques for improving retrieval performance. Relevance feedback is the process of iteratively modifying an original query from relevant/non-relevant documents (Spink and Losee 1996). Not only did the IR community focus attention on the development of different techniques for improving retrieval performance using relevance feedback and query expansion, IR researchers also focused on the evaluation of the effectiveness of these two approaches early on. Salton and Buckley (1990) have laid out an evaluation framework, which is used to assess and compare techniques using relevance feedback and query expansion. Two important issues raised by the paper are the use of a residual collection and the computation of average precision at three specific recall points of 0.25, 0.5 and 0.75 which is called three point fixed recall. A residual collection is created by removing all documents previously seen by a user regardless of whether they are relevant or not from the original collection. The evaluation process is then done using the reduced collection. The main idea for using the residual collection is to assure that we assess a technique based on *new* information retrieved. However, this technique has bias towards queries that have more relevant documents in the test collections and performs poorly in the first initial iterations (Ruthven and Lamas 2003). The average precision at particular fixed recall points offers an opportunity for easy comparison among different techniques. The best technique found in this evaluation (Salton and Buckley 1990) over all test collections is Ide dec-hi which is still very competitive, as pointed out by recent studies (Lopez-Pujalte et al. 2003, Drucker et al. 2003).

In the UM community, most of the studies focusing on retrieval performance uses two commonly used metrics in IR: precision and recall (Salton and McGill 1983). These standard metrics are used to evaluate the accuracy of a system running with and without the user model over a controlled group of users and/or with a group of real world users (Billsus et al. 1999, Bueno et al 2001 and Magnini et al. 2001). These studies use their own test collections and metrics and thus make it difficult to compare with each other.

Our approach

We capture user intent by analyzing the retrieved relevant documents and use this information to modify the user query proactively. We partition user intent into three formative components: Context, Interests and Preferences. The context provides insight into the user's knowledge and represents the connections among a user's goals for easy explanation. It is captured in Context network (C). C consists of concept nodes and relation nodes. A concept node is a noun phrase. We have the two kinds of relation nodes: set-subset (denoted as ''isa'') and related to (denoted as ''related_to"). C is created dynamically by finding the set of common subgraphs in the intersection of retrieved relevant documents. Figure 1 (a) shows an example of a context network of an analyst who is searching for information on terrorism and suspicious banking transactions. The Interests capture the focus and direction of the individual's attention. It is captured in the interest set (I). Each element of *I* consists of interest concept (a) and interest level (L(a)). An interest concept represents the concept that an analyst is currently focusing on while an interest level is any real number from 0 to 1 representing how much emphasis he places on this particular concept. *I* is initially determined from the current query, and the set of common subgraph. Figure 1(b) shows the example of an interest set of the above analyst. Lastly, the Preferences describe the actions needed to perform to achieve the goals. We capture Preferences in a Bayesian network (Jensen 1988) which consists of three kinds of nodes: pre-condition, goals and action nodes. Precondition nodes represent the requirements to achieve the goal nodes. Goal nodes represent the tools that are used to modify a user's query. We currently have the two tools: filter which narrows down a query semantically and expander which broadens up a query semantically. Figure 1 (c) contains the preference network for the above analyst. Now, suppose that the analyst is interested in finding the latest information of suspicious banking transaction. He issues the query "Banking transaction" which will be converted into a query graph (as shown in Figure 2 (a)). The system will use the information on his context, interest and preference to modify this query by making the concepts of the original query more specific which reflects his intentions of finding the relationships

between suspicious banking transactions and terrorism activities, as shown in Figure (b). Instead of asking for general information on banking transaction, he asks for specific deposits, withdraw activities which relate to money laundering activity. For more detail about our approach, please see (Santos et al. 2001, 2003a, 2003b)

Figure 1: Examples of context network, interest set and preference network

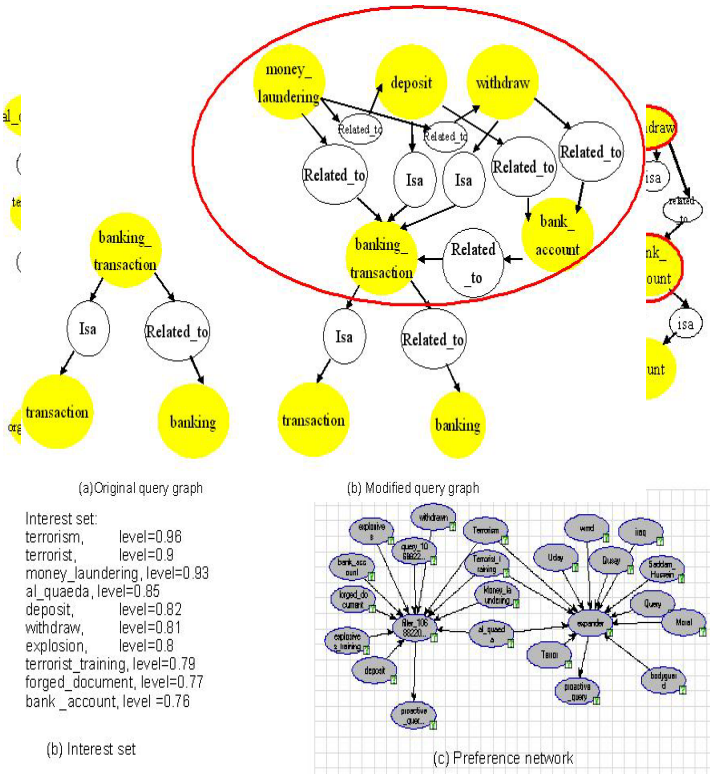

Figure 2: Example of original query graph modified query graph

Experiment setup

Testbed

We choose CRANFIELD as our first testbed because it is one of the oldest test collections in the IR community and is widely used for evaluating the effectiveness of any relevance feedback and query expansion techniques (Salton and Buckley 1990, Lopez-Pujalte et al. 2003, Drucker et al. 2003). It contains 1398 papers on aerodynamics and 225 queries with relevancy assessment. A relevant document with regards to a specific query is rated on a numerical scale from 1 (most valuable) to 4 (least valuable). We evaluated using a set of 43 queries with the property that the total relevant documents, retrieved by both approaches, in the top 15 is at least 2 and at least 6 relevant documents are not in the top 15. We choose this set of queries because evaluating using residual collection has bias towards queries that have more relevant documents.

Procedure for TFIDF approach

We re-implement TFIDF and Ide dec-hi and run Ide dec-hi with TFIDF as follows: Each query in the testbed is converted to a query vector. The query vector will be compared with each document vector in the collection. We take the first 15 returned documents for relevancy assessment. We use Ide dec-hi to expand the original query (Salton and Buckley 1990). For each query in the collection, we run the system twice. We refer to the first run as *initial run* and the second run as *feedback run*. For each query, we compute average precision at three point fixed recall (0.25, 0.5 and 0.75).

Procedure for our approach

We evaluate the effectiveness of our user model through a set of four experiments, which assesses the user model's short-term effects on a single query, and the user model's long-term effects on the whole search session. There are three main reasons for our experimental design. First, in our approach, the relevance feedback for one query can be used for the same query immediately or related queries later on. Therefore, we would like to assess the immediate effects of the model obtained by modifying the same query. Also we want to assess the effects that are obtained by using user model to modify any queries during a whole search session. Second, we would like to see where our approach stands if we follow the same procedure designed for the Ide dec-hi approach. Lastly, we would like to test our hypothesis that the retrieval performance should be improved if our user model has some knowledge about a user.

We divide our evaluation into two groups of experiments. The first group starts with an empty user model and the second group starts with the seed user model with the interest relevancy set being reset to empty. The seed user model is the user model that is generated after we run the system through the set of queries once. Our goals for this procedure are to evaluate:

(i) How effective the user model is in terms of improving retrieval performance.

(ii) How competitive the improvement obtained by using our user model compares with the improvement otained by using Ide dec-hi with TFIDF.

Starting from knowing nothing about a user or the search domain, the first group of experiments includes:

Experiment 1: For each query in the testbed, we run our system with an empty user model, give feedback and rerun the system with the same query.

Experiment 2: We also start this experiment with an empty user model. However, for each query, we update our user model based on relevance feedback.

At the end of this experiment, we save the final user model and use this as the seed user model for our next group of experiments. The second group of experiments is similar to the first one except that we start with the seed user model.

For each query in the testbed, a query graph is constructed. We use Link Parser (Sleator and Temperley 1993) to parse a natural language query. Link Parser comes with a limited dictionary which sometimes produces incorrect parse trees for the sentences consisting of words that are not contained in its dictionary. Therefore, out of 43 queries, 22 query graphs are manually created and loaded into the system to make sure that we have correct query graphs to work with. The document graph is generated from natural language for each document in the database without modification. Please see (Santos et al. 2001) for more detail.

Analysis of results

The precision of the initial run and feedback run using residual collection of TFIDF and Ide dec hi are 0.11 and 0.21 respectively, and they are 0.19 and 0.31 when evaluated using original collection. Our results using residual collection are similar to those in previous publications (Lopez-Pujalte et al. 2003). Our goals initially set for these four experiments have been met. Experiments 2 to 4 show that by using our user model, the precision of the feedback run is always higher using residual and original collections compared to the initial run. Among four experiments, we can see that Experiment 4 performed competitively well compared to Ide dec-hi in the feedback run while it offers the advantages of having higher precision in the initial run compared to TFIDF.

Experiment	Run	Residual	Original
Exp 1	Initial run	0.12	0.29
	Feedback run	0.14	0.33
Exp 2	Initial run	0.12	0.28
	Feedback run	0.14	0.31
Exp 3	Initial run	0.13	0.30
	Feedback run	0.19	0.33
Exp 4	Initial run	0.12	0.29
	Feedback run	0.17	0.33

Table 1: Average precision at three point fixed recall.

Moreover, we found out that the number of relevant documents in the top 15 of the initial runs in all four experiments using our approach for all queries in the testbed is always higher than that of the initial runs using TFIDF (as shown in the last row of Table 2). As the CRANFIELD collection has classified relevancy of a document on a 4-point numeric scale for all relevant documents, we investigated further to see how many good relevant documents our approach has retrieved from the initial runs compared to TFIDF. Table 2 shows that all of our four experiments retrieved more documents ranked 1 than the TFIDF approach. We also retrieve more relevant documents in all ranks in top 15.

Significance of the results from a human factors perspective.

There are three important implications from our results which can be used for designing a more user friendly and effective IR application. The ultimate goal of a user in using an IR application is to retrieve as many as possible quality documents relevant to his/her information needs (Baeza-Yates and Ribeiro-Neto 1999). Our results show that we have retrieved more good quality documents in the first initial interactions than TFIDF approach in all experiments. This implies that the amount of time required to complete an information seeking task will be reduced and therefore a user's objective efficiency will be improved, if we use the metrics in our pilot usability testing (Santos et al. 2003a). This opens a window of opportunity for using our user model approach to help users to satisfy their information needs quickly. In addition, our seed user model actually helps improve retrieval performance in Experiments 3 and 4. The seed user model contains the knowledge that a user has learned in the searching process. It implies that the retrieval performance is improved if our user model has some knowledge about a user and searching domains.

Relevancy	TFIDF	Exp	Exp	Exp	Exp

		1	2	3	4
1	13	19	19	20	22
2	33	41	42	41	39
3	49	64	62	65	62
4	26	31	32	31	28
Total	121	155	155	157	151

Table 2: Total number of top 15 retrieved relevant documents.

It also implies that we can use a seed model of an expert user to help novice users to find information quickly as well as use a seed model to better adapt to a user's searching styles and strategies. In all four experiments, the retrieval performance of the system after receiving feedback from users is higher compared to the first initial interactions. It means that the knowledge captured through interactions between a user and the target system helps improved retrieval performance. In the long-run, this can be explored further to improve the interactivity between a user and an IR system.

Conclusion

The results of this evaluation show that by exploring user intent, we achieve competitive performance in the feedback run compared to Ide dec-hi; at the same time our user model approach offers the advantages of retrieving more quality documents at the initial run compared to the traditional TF-IDF approach. The results can be used by designers of information retrieval systems to reduce the iterations that a user has gone through to retrieve good relevant documents and thus improves user satisfaction. Our approach is different from the Ide dec-hi approach in that the knowledge of user intent learned over time by our user model can be used for future retrieval process. We also show that the more knowledge learned about users and searching domains, the better a target information retrieval application performs. These results imply that our user modeling approach can be used to improve and reinforce learnability, efficiency and interactivity between a user and an IR system. We would like to extend our empirical evaluation and further validate our hypothesis on larger testbeds and large scale usability testing.

Acknowledgement

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

References

Baeza-Yates R. and Ribeiro-Neto B. 1999. *Modern Information Retrieval*. Addison Wesley Longman

Billsus D. and Pazzani M. P. 2000. User modeling for adaptive news access. *Journal of User Modeling and User-Adapted Interaction.*Vol 10 (2/3). Pages 147-180.

Bueno D. and David A. A. 2001. METIORE: A personalized information retrieval system. *In Bauer, M., Vassileva, J. and Gmytrasiewicz, P. (Eds.). User Modeling Eight International Conference, UM2001.* Pages 168-177. Berlin, Springer.

Brajnik G., Guida G and Tasso C. 1987. User modeling in intelligent information retrieval. *Information Processing and Management.* Vol 23(4). Pages 305-320.

Cleverdon C. 1967. The Cranfield test of index language devices. *Reprinted in Reading in Information Retrieval Eds. 1998*. Pages 47-59.

Drucker H. and Shahrary B. and Gibbon C. 2002. Support vector machines: relevance feedback and information retrieval. *Information Processing and Management.* Vol 38(3). Pages 305-323.

Loper-Pujalte C., Guerrero-Bote B. and De Moya-Anegon F. 2003. Genetic algorithms in relevance feedback: a second test and new contributions. *Information Processing and Management.* Vol 39(5). Pages 669-697.

Magnini B. and Strapparava C. 2001. Improving user modeling with content-based techniques. *In Bauer, M, Vassileva, J, and Gmytrasiewicz, P. (Eds). User Modeling Eighth International Conference, UM2001.* Pages 74-83. Berlin, Springer.

Nguyen, H. 2004. Capturing User Intent forInformation Retrieval. In *AAAI 2004 Doctoral Consortium.* To appear

Salton G. and Buckley C. 1990. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science.* Vol 41(4), 288-297.

Salton G. and McGill M. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company.

Santos E. Jr., H. Nguyen, Q. Zhao and E. Pukinskis. 2003a. Empirical Evaluation of Adaptive User Modeling in a Medical Information Retrieval Application. In *Proceedings of the ninth User Modeling Conference UM2003,* Johnstown. Pennsylvania. Pages 292-296

Santos E. Jr., H. Nguyen, Q. Zhao and W. Hua. 2003b. User Modelling for Intent Prediction in Information Analysis. In *Proceedings of the 47th Annual Meeting for the Human Factors and Ergonomincs Society (HFES-03).* Pages 1034-1038.

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.* Pages 412-423.

Sleator D. D. and Temperley D. 1993. Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies.* Pages 277-292

Spink and R. M. Losee. 1996. Feedback in Information Retrieval. In Williams, M. Eds. *Annual Review of Information Science and Technology*. Vol 31.Pages 33-78.

Ruthven and M. Lalmas. 2003. A survey on the use of relevance feedback for information access systems, *Knowledge Engineering Review*. Vol 18(1).