

# Using User's Context for IR Personalization

N.J. Belkin, G. Muresan, X.-M. Zhang  
School of Communication, Information and Library Studies  
Rutgers University, New Brunswick, NJ 08901  
+1-732-932-7500  
[belkin, muresan, xzhang]@scils.rutgers.edu

## 1 INTRODUCTION

We believe that knowledge of the user's context is vital to personalizing information retrieval (IR) interaction, and furthermore, that such knowledge is best obtained through implicit sources of evidence, e.g. inferences made on the basis of the user's past or current behaviors. We further believe that it is now well past time to *test* whether such knowledge really does affect the interactive IR experience. Some contextual factors that have been suggested as being important to consider in improving IR performance, by us and others, include:

- searcher's familiarity with or knowledge of the topic;
- searcher's experience of searching for information;
- documents which the searcher has previously found (un)useful;
- genre of desired documents;
- purpose of the search (use to which retrieved documents would be put);
- task which led the searcher to information seeking;
- what else the user is doing at the time of information seeking.

In order to test our ideas about what aspects of a user's context might be important in this sense, and how knowledge of these characteristics could be utilized to affect various IR techniques, our group at Rutgers University participates in the TREC HARD track. Here, we present an overview of the general HARD approach, our position on how this issue should be addressed, and how we have attempted, and are attempting to implement such knowledge.

## 2 THE HARD TRACK

The HARD Track investigates the effect of knowledge of user's context on IR system performance in the following way. Search topics are constructed by assessors, with respect to issues of interest to them. These topics follow the general

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '04, July 25-29, 2004, Sheffield, UK.

Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00

TREC format, with the addition of categories of metadata whose values describe various aspects of the assessor's context. In TREC 2003, the metadata were:

- familiarity with the topic
- desired genre of retrieved documents
- purpose of the search
- specification of geographic focus of documents

In TREC 2004, the categories of metadata were reduced to:

- knowledge of the topic
- desired genre of retrieved documents
- documents should be about USA, or not USA

In addition to these categories of metadata, assessors also specify one or two documents related to the topic, and the granularity of response they require.

Participating sites are initially given the document corpus and a set of training topics, each with the metadata and 100 documents which have been judged either *not* relevant, *soft* relevant (i.e. on topic), or *hard* relevant (i.e. on topic, and satisfying the metadata conditions). Then, the set of test topics is distributed, *without* the user metadata. Each site constructs a query for each topic, searches the corpus and returns a ranked list of documents for each topic. This constitutes the *baseline* run. Then, the metadata and other information for each topic are distributed to all sites. The sites then use information derived from the metadata to modify the retrieval techniques (e.g. modify the query, re-rank the baseline list). In addition, they may submit a clarification form to the assessor, asking one simple, limited question concerning some aspect of the initial retrieval performance (e.g. which of these clusters of retrieved documents do you find most interesting). The sites then submit one or more new runs, based on the information received. These are the *test* runs.

The results of baseline and test runs are pooled, and evaluated by the original assessors according to the three categories of relevance. The test of the utility of the modifications that have been made is the difference in performance between the baseline and test runs, *judged according to hard relevance*. For a detailed explanation of the TREC 2003 HARD track, see Allan (2004).

## 3 RUTGERS' APPROACH TO HARD

### 3.1 Introduction

The goal of our work in the HARD track is to test techniques for using knowledge about various aspects of the information seeker's context to improve IR system performance. We are particularly concerned with such knowledge which could be

gained through implicit sources of evidence, rather than explicit questioning of the information seeker. We therefore do not submit any clarification form. Of the categories of metadata and related information which are available in the HARD track, we have chosen to investigate:

- searcher familiarity with (or knowledge of) the topic, since there is some evidence that it is important (Kelly and Cool, 2002) and evidence of this could be gained through observation of previous behavior;
- desired genre, for the same reasons (e.g. Rauber & Müller-Kögler, 2001); and
- related texts, since they could be inferred on the basis of both past and current behaviors.

### 3.2 HARD 2003

In TREC 2003 we attempted to test the following hypotheses:

**H1:** People familiar with a topic will want to see documents which are detailed and terminologically specific; people unfamiliar with a topic will want to see general and relatively simple documents. This we operationalized by promoting the value of documents which scored toward the unreadable end of readability scales for people highly familiar with the topic, and by promoting the value of documents which scored toward the easily readable end of the scales for people unfamiliar with the topic.

**H2:** Different document genres can be identified by their vocabularies. This we operationalized by constructing language models for all the retrieved and for just the hard relevant documents for each training topic. By comparing language models, we then identified words which occurred with greater than expected probability in the relevant documents, for all topics which had the same genre. These words were considered indicators of the genre and were used for query expansion for topics which requested that genre.

**H3:** Certain document sources will be relevant, or not, to different desired genres. This we operationalized by promoting or demoting the score of documents, or by removing documents from the ranked list, according to their source and the requested genre.

**H4:** If there are texts which the information searcher has identified as related to the topic, using them as the basis for automatic query expansion will improve retrieval performance. This was operationalized by choosing terms for query expansion from the related texts.

We understood that there are, in general, two ways in which to take account of the metadata. One is to modify the initial query from the (presumed) searcher, before submitting it for search; the other is to search with the initial query, and then modify (i.e. re-rank) the results before showing them to the searcher. We used both of these techniques in taking account of the different types of metadata.

Our results in the 2003 HARD Track indicated a few interesting trends, but were generally poor. Detailed analysis suggests that we did gain some advantage from using the metadata to modify the baseline queries, in some respects, and query expansion via related documents did help. But the ways in which we used the metadata to modify rankings and

queries were quite ad hoc, and without real theoretical justification, which could go some way toward explaining negative results. A more detailed report on our HARD 2003 experience can be found in Belkin et al. (2004).

There were some significant problems with HARD 2003. The training data were insufficient, the familiarity scale did not actually judge the assessor's real familiarity with the topic, and there was insufficient representation of different values of the different metadata for training and testing purposes. Thus, for HARD 2004, both the number of metadata factors, and the number of values which they could take, were reduced.

### 3.3 HARD 2004

In HARD 2004, we are attempting to make our hypotheses more formal, and to move from ad hoc implementations of our hypotheses to more principled ones. This is still in progress, but we outline them here. In addition to query expansion from related documents, Rutgers is investigating the following two issues:

- how can we take account of a searcher's knowledge of a topic to improve retrieval performance; and
- how can we take account of knowledge of desired genre to improve retrieval performance.

In HARD 2004, there are only two values of knowledge of a topic: *little*; and *a great deal*, and there are only three values of genre: *news-report*; *opinion*; and *other* (the corpus consists of news sources). With respect to these issues, we consider the following hypotheses:

**H1:** People with a great deal of knowledge of a topic will want to see documents which are detailed and terminologically specific; people with little knowledge of a topic will want to see general and relatively simple documents. This is the same hypothesis that we had with respect to familiarity in TREC 2003. However, we are investigating the use of different readability measures, which are more directly concerned with terminology than those we used in TREC 2003.

In addition, we are investigating two new hypotheses with respect to knowledge of the topic. These have to do with findings that people with little knowledge of a topic cannot interpret and understand *abstract* concepts in the topic domain as well as those who have good knowledge, and that words indicating *concrete* concepts are in general more easily understood than abstract ones. This leads us to the following:

**H2:** People with little knowledge of the topic will prefer documents with a low ratio of abstract words to total words, and of abstract words to concrete words. People with good knowledge of a topic will prefer documents which have a high ratio of abstract words to total words, and of abstract words to concrete words. This hypothesis leads to a re-ranking strategy.

**H3:** Adding concrete terms to the initial query from the topic domain (as determined by initial retrieval results) will lead to more effective results for people with little knowledge of the topic; adding abstract terms from the topic domain will lead to more effective results for people with a great deal of knowledge of the topic. This is a query modification strategy.

With respect to genre, we have several hypotheses, one of which is directly related to those of TREC 2003. They are all based on the idea that no matter what the topic, documents of specific genres will share some common characteristics which can be identified through different sorts of analyses.

**H4:** The differences between the genres of news-report and opinion can be identified according to the degree of subjectivity or objectivity of a document, as determined by various linguistic features of the documents (cf. Rittman, 2004). This leads to a classification and re-ranking strategy.

**H5:** Different document genres will have different characteristic vocabularies, regardless of topic. This is essentially the same hypothesis as for TREC 2003, and we investigate it by again developing language models for the topic in general (i.e. soft relevant), and those for the different genres within each topic. Words which occur with greater than expected frequency with respect to the topic models for a particular genre, across all topics, will be indicative of the genre's vocabulary. This technique can be used both to identify words which can be added to a query (query modification strategy), and to classify documents which belong to a specific genre (re-ranking strategy).

**H6:** Different document genres will have different discourse-level features characteristic of each genre, regardless of topic. We will determine these features with the training collection, and use them to classify initially retrieved documents. This leads to a re-ranking strategy.

## 4 CONCLUSION

We have outlined a general experimental approach to evaluating the effectiveness of taking account of different aspects of a searcher's context in personalizing retrieval to that person. Although initial results are still very sketchy, we believe that following this route is necessary in order to determine whether context really is important, what aspects of context are really important, and how knowledge of such aspects can best be taken account of.

## 5 REFERENCES

- Allan, J (2004) Overview of the TREC 2003 HARD track. In Proceedings of the 12<sup>th</sup> Text REtrieval Conferenc (TREC 2003). Retrieved from [http://trec.nist.gov/pubs/trec12/t12\\_track.index.html](http://trec.nist.gov/pubs/trec12/t12_track.index.html) on 28 May 2004.
- N.J. Belkin, D. Kelly, H.-J. Lee, Y.-L. Li, G. Muresan, M.-C. Tang, X.-J. Yuan, X.-M. Zhang (2004) Rutgers' HARD and Web Interactive Track Experiences at TREC 2003. In Proceedings of the 12<sup>th</sup> Text REtrieval Conferenc (TREC 2003). Retrieved from [http://trec.nist.gov/pubs/trec12/t12\\_track.index.html](http://trec.nist.gov/pubs/trec12/t12_track.index.html) on 28 May 2004.
- Kelly, D. & Cool, C. (2002) Effects of topic familiarity on information search behavior. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries – JCDL 2002* (pp. 74-75). New York: ACM.
- Rauber, A. & Müller-Kögler, A. (2001) Integrating automatic genre analysis into digital libraries. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL 2001* (pp. 1-10). New York: ACM.
- Rittman, R. (2004) Adjectives as Indicators of Subjectivity in Documents. In Proceedings of the 2004 Annual Meeting of the American Society for Information Science and Technology. in press.