HARD Experiment at Maryland: from Need Negotiation to Automated HARD Process

Daqing He
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland, 20742
daqing@umiacs.umd.edu

Dina Demner-Fushman
Department of Computer Science
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742
demner@cs.umd.edu

Abstract

Our aim of participating in this year's High Accuracy Retrieval from Documents (HARD) track is to explore the possibility of developing an automated HARD retrieval model by leveraging existing models and theories about information need negotiation in information science. The clarification questions we developed are related to four different aspects of search topic, and four different techniques were developed to fully use the information collected from the user through these questions. Our initial analysis of the results indicates that this is a promising approach.

1 Introduction

Searching for information is increasingly common in people's life. Modern techniques based on "free text" indexing and ranked retrieval have proven to be scalable and robust. Batch mode information retrieval (IR), which essentially studies retrieval algorithms, receives a great deal of attention. Significant improvements have been achieved both in academic and commercial paradigms. Many people associate the improvements, especially those achieved in academic paradigm, with controlled experiment frameworks, such as Text Retrieval Conference (TREC).

However, the initiative of searching for information ultimately lies with human. It is people who pose the questions, interpret what they read, and determine when their needs have been met. Especially in modern retrieval process, end users, who are not necessarily search experts, nor domain experts, leverage easy access to full text to support increasingly focused exploratory searches via iterative refinement (Marchionini, 1995). Therefore, the ultimate goal of retrieval systems is not to generate the best possible ranked list for a given search query, but to provide the best information access mechanisms to users so that they can easily find needed information, and have a pleasant search experience.

Based on this view of retrieval process, many researchers concentrated on interactions in information retrieval process, and designed experiments within a controlled experiment framework. The interactive track in TREC was an effort dedicated to this task. Many interesting research results have been achieved via this approach, but many of its limitations have also been shown. It is widely accepted that interactive IR experiments are difficult to design, expensive to conduct, limited in their small scales, and hard to compare cross-site. Our past experience with interactive IR, especially experiments we conducted for interactive track of Cross Language Evaluation Forum (iCLEF) (He et al., 2002; Dorr et al., 2003b), provides us with some first-hand knowledge of these limitations.

We see HARD, a new track of TREC 2003, as an opportunity to ask interesting questions about the real human retrieval process, especially the interactions between human and the retrieval process, and at the same time, to design a relatively easy, cheap, and large scale controlled experiment to find answers to our questions, and to compare the results to these of other sites.

To us, HARD experiment models the retrieval

process differently to both batch and interactive IR. For better representation of the actual retrieval process, HARD allows interactions between the users and the retrieval system, which is like interactive IR. However, to avoid the difficulty of managing full interactions, HARD only allows one iteration of interactions. The interaction is conducted by letting the system generate a set of clarification questions to be answered by the user. Then the system uses the answers to improve its search effectiveness. In addition, HARD uses measures on returned rank list as indicators of the performance rather than measures on relevance judgments often used in interactive IR.

HARD experiment, to some extent, can be viewed as a simplified model of information need negotiation services, which is a well studied area in information and library science (Taylor, 1968). The process of generating clarification questions in HARD experiment is a simplified version of information need negotiation, or reference interview. Once we achieve this transformation of models, we will have opened a rich resource for us to borrow. Therefore, our approach in this year's HARD experiment is to leverage existing theories, models, ideas, and resources in information need negotiation to design and implement an automated process of generating clarification questions and utilizing their answers to improve the ranked list of documents for a given query statement.

In the rest of the paper, we are going to briefly introduce the idea of information need negotiation in information science, and then move to present our approach of leveraging existing models to our automated HARD process. In the remaining sections, we discuss some preliminary analysis of our experimental results, and finally we conclude with some indications to future directions.

2 Information Need Negotiation

Information need negotiation is a reference interview in the library setting. It is a communication between an information specialist and a user, in which the users present their information requirements, and the specialist clarifies these needs to develop an appropriate, mutually agreeable search strategy. In his classic paper about information need negotiation (Taylor, 1968), Taylor identifies this process as question negotiation, since he believes that the query issued by a user is not a command, but a question that the user wants to be answered by the information specialist. Because a user in IR tries to search for something that he/she does not know, the negotiation process contains complex actions where two persons interact to achieve the goal of identifying the need and find an appropriate search strategy. This is why the negotiation of reference questions is one of the most complex acts of human communication.

Based on studies of librarians and information specialists, Taylor identified the following five general types of information often occurring during an information need negotiation process:

1. determination of the subject that the user is searching on;
2. objective and motivation for the current search;
3. personal characteristics of the user;
4. relationship between the search statement and the file organization in the collection; and
5. anticipated or acceptable answers.

With the advancement of information technology, not all information need negotiations are conducted face to face between a user and an information specialist. One of the examples of remote need negotiation services is conducting the need negotiation via Email. Abels conducted a three-phased project at University of Maryland to explore the email negotiation process (Abels, 1996). She identified five approaches often used in email negotiation process: (1) piecemeal, (2) feedback, (3) bombardment, (4) assumption, and (5) systematic. Her analyses showed that the systematic approach yielded successful need negotiation, and it was clearly the most efficient in terms of number of messages needed in the process. In the systematic approach, the specialists responded to a request with a list of questions covering all related aspects. The questions were organized in a logical way and included both open- and closed-end questions. At the later stage of the project, Abels designed a request form to include all the questions that would be asked in the systematic approach of need negotiation. The essential content of the form are questions about the personal data of the user,

subject to be searched, and preference/constraints on the search results.

3 Constructing automated HARD process

Our HARD process naturally divided into two stages. The first one was to automatically generate a set of clarification questions to probe for more information about the topic, the person who had the need, and his/her preferences regarding search results. The second stage was to automatically utilize the answers to questions in the clarification forms. The design of the questions in the clarification forms and the methods utilizing the answers to the questions are closely related. The questions were selected based on our understanding of how the answers will be applied to improve the retrieval effectiveness. The automated process for utilizing these answers was designed to include as much as possible of the information from the answers.

3.1 Generating Clarification Questions

We combined Taylor's question negotiation model and Abels' email reference forms, and designed our own set of clarification question types. The clarification questions came from four aspects of context information related to a given query. They are presented in each of the following sub-sections, respectively.

3.1.1 Users' preference to sub-topic areas

Documents that are retrieved for a given query can be classified into multiple sub-topic areas. One reason is that the search topic naturally has multiple facets. Another reason could be that the query terms have multiple senses, which results in the search system retrieving documents related to several sub-topics. For example, topic 87 is about *Egyptian cotton*, and its top ranked documents retrieved by our baseline system covered sub-topics ranging from an advertisement for Macy's white sale to the history of Egypt, to cotton leaf worm or child labor in Egypt's cotton industry.

Although there could be cases when the user is interested in more than one sub-topic in the same search, prevalently the user is interested in only one of them. Differentiating the intended sub-topic area from those irrelevant ones would help the system to avoid placing irrelevant documents at the top of the ranked list. Therefore, the first aspect of the clarification questions we tried was to probe the user's preference to the sub-topic areas.

Two tasks were needed to enable us to generate questions for the user to select among sub-topics. The first one was to identify the sub-topic areas existing in the top ranked baseline retrieval results. The second task was to present the user with a concise description for each sub-topic area so that the user could select.

For the first task, we used automatic clustering method. Clustering has been shown as an effective method for organizing and presenting closely related documents (Hearst and Pedersen, 1996). During clustering, we only used top ten retrieved documents for each query because our training results demonstrated that this part of retrieval results contains little noise. In addition, space limitation imposed on clarification forms by HARD track guideline meant that only a small number of clusters (i.e., less than 5 clusters) were possible. Restricting the document pool size also helped to obtain tight small-sized clusters of closely related documents.

We clustered the documents using the Lighthouse implementation of the Ward's hierarchical clustering algorithm (Leuski and Allan, 2000). The distance measure used in the clustering was the cosine value of the angle between the vector representations of two documents. The weight for each term in the vector representation is defined as the product of the term's frequency in the document and its inverted document frequency. The average number of clusters for a topic was 3.5.

For the task of presenting a concise description of clusters, we explored three different approaches. We first tried to use top 10 highly weighted terms in the cluster as the cluster representative, but were not satisfied with the outcome when we tried it on the training data.

We then tried another approach based on the genre of the documents. Most of the documents in the HARD collection are news articles. It has been shown that titles and first sentences of news domain articles contain enough information to represent the main topic of the document (Dorr et al., 2003a). Our second approach, therefore, was to select the titles or the first sentences, in the absence of the titles, from each document in a cluster as candidates for

the cluster representative. These candidates were ranked based on the normalized sum of the weights of the terms in them, and the one at the top was selected as the cluster representative. An example of the cluster representative generated by this method is *"Middle East Economic Briefs"*.

The third approach we tried was to use a multi-document headline generation tool called GOSP. It was developed at Information Science Institute, University of Southern California (Zhou and Hovy, 2003). Before the documents could be processed GOSP, we tokenized and POS-tagged the documents using MXPOST and MXTERMINATOR software (Ratnaparkhi, 1996). An example of the cluster representative generated by this method is *"MIDDLE EAST COUNTRIES BEIRUT LEBANONS NATIONAL/1998 LEBANESE MINISTER OF PUBLIC WORKS AND TRANSPORTATION NAJIB MIKATI/CENTRAL BANK BE PRIVATIZED"*.

3.1.2 Users' recent experience with the search topic

The second type of questions in our clarification forms was about user's characteristics. In this year's HARD experiment, we concentrated on the user's recent experience with searching on the subject area. This information is important because the user's information need on a topic could be evolving over time, and the answer to this question can help us determine the current status of the user's need on the topic. In addition, the answers to this question provide the necessary information to perform query expansion when the user does not select any clusters as relevant.

The question directly asks for the terms related to the user's recent search on the topic. Our question starts with an inquiry about whether or not the user had seen any relevant documents recently. If the answer is positive, we ask for the key words that would best describe the document. In the question about the key words, we specifically ask the user to provide highly representative content words, person or organization names, and terms related to locations.

3.1.3 Users' preference to sub-collections

The HARD collection contains news articles and US government documents. Among news articles, there are documents from news agencies inside US, and those from Xinhua news agency. A user who has particular information need usually does not necessarily have the same preference for the documents from different sources. For example, a user who is interested in international response to an event will be more willing to read articles from Xinhua news agency than US government documents. Therefore, knowing these preferences would help the retrieval effectiveness since the retrieval system can pull the documents from the preferred sources to higher ranks. However, these preferences are usually user dependent and topic dependent, so asking the user to provide such information is much easier and effective than letting the system automatically infer user preferences.

Our question about the user's preference to sub-collections is designed to facilitate quick response from the user. Users were asked to rank their preferences on a scale from 1 to 5, with 1 as the most preferred one.

3.1.4 Users' anticipation of result formats

Although retrieval systems usually return documents as the default format of the results, different users under different information needs may prefer different formats, such as documents, passages, sentences, or even straight answers. For example, this year's HARD experiment guidelines contain a specification to identify the user's preferred result format. Our question to this type of information was designed to be a straight selection of the format the user wants for this particular search topic.

3.2 Applying Answers to Clarification Forms questions

Our automated process utilized the information obtained from the answers to the clarification forms in three ways, namely, term extraction for query expansion, preference extraction for document re-ranking in a ranked list, and evidence combination for ranked lists merging. To test the effectiveness of these three ways in isolation or combination, we designed several runs to include either only one of them or combinations of them. In the remainder of this section, we are going to present our methods for query expansion, document re-ranking and ranked list combination.

3.2.1 Query Expansion

The information used in query expansion is from two sources. The first source is a set of texts that includes the description and narrative sections of the topic statement (since we only used the title part of each topic statement in our baseline run); the documents belonging to the preferred clusters marked by the user; and the relevant documents provided in the meta data part, if meta data were used in the run. Because texts from different sources were rather different, we extracted terms from them separately, and combined the terms only when using them to expand the query. Terms were extracted based on the combination of their term frequencies in the documents and their inverted document frequencies.

The second source is the set of terms provided by the user when he/she answered the clarification questions about their recent experience with the search topic. Since the answers to these questions were already a set of terms, and they were provided directly by the user, we did not perform any further term extraction before expanding the query using these terms.

The expanded queries were constructed by including the terms from the original baseline query, and the terms from the two sources above. Besides the weights calculated during the term extraction (i.e., the weights associated with the terms from the first source), we also assigned a predefined weight to each term based on its origin. We assigned equal weight (i.e., 20) to terms from the original query, and those from the second source above. The rationale is that both sets of terms were directly provided by the user. We then normalized the weights of the terms from the first source so that the highest weight among them is only half the weight of the terms from the original query (i.e., the weight is 10), and all the other weights from the first source were mapped proportionally. This reflected our view that we trust more the terms directly issued by the user, and less the terms we extracted. If a term appeared in multiple sources, the weights of its several appearances were combined.

3.2.2 Document re-ranking

Document re-ranking in our approach referred to boosting or suppressing documents with certain features based on the user's preference. The goal of this method was to improve the retrieval effectiveness by rearranging the ranks of documents.

The information used in helping us re-rank the documents included the answers about the user's preference to sub-collections, and the answers about the user's preference to a time period covered by the HARD collection. When the meta data were included in the refined run, we also used the information about the user's preference to the genre of the documents to help us in re-ranking.

Two approaches can be applied in document re-ranking. There could be aggressive re-ranking, in which the documents possessing certain features are boosted or suppressed to the maximum. For example, if we know a user prefers documents from one sub-collection, aggressive re-ranking would put all the retrieved documents from this sub-collection at the top of the ranked list. This approach sometimes makes sense. For example, if we knew that the user is really interested only in government documents, we could achieve best results putting all retrieved government documents at the top of the ranked list.

The other approach is conservative re-ranking, that is, boosting or suppressing some documents only to some degree. For example, using a small predefined boost or suppress factor to perform re-ranking. This approach is appropriate when there is not much training information to be used in the development, which means that we could not give too much trust to the re-ranking algorithm.

Our re-ranking algorithm in this year's HARD experiment was the mixture of aggressive and conservative approaches. On the one hand, there seemed to be a clear indication of user's preference when the user marked that he/she wanted certain genre of documents (e.g., government documents), which encouraged us to use aggressive re-ranking approach. But on the other hand, there were not much training data for us to really test our re-ranking algorithm, which indicated the conservative re-ranking approach to be more appropriate. At the end, we adopted the aggressive re-ranking algorithm when we used the genre meta data, but kept conservative approach in re-ranking when we used other information.

3.2.3 Ranked list combination

People have demonstrated that, if designed carefully, merging ranked lists from different resources could improve the retrieval effectiveness (Kamps et al., 2002). In addition, it is also a safer approach to merge a ranked list that we do not know much about with the ranked list that we trust to certain extent. In this year's HARD experiment, we performed ranked list combination for both of the reasons above.

We adopted a linear combination approach where the scores of documents in two lists were first normalized proportionally to the highest score in the list, and then combined linearly by applying a predefined list-specific weight factor λ. During the training stage, we noticed that we can achieve even better performance if we micro-adjust the list weight λ with the difference between the highest scores of the lists.

We combined the ranked list of results of one of the refined runs with our baseline run. The chosen refined run utilized our query expansion method and document re-ranking based on the answers from clarification forms and meta data.

3.2.4 Passage and sentence retrieval

Our passage retrieval module assumes that the relevance of a passage is related to the frequency of the query terms in the passage, the importance of these query terms (i.e. their weights), and the relevance of the document that contains the passage to the query. Among the three factors, we gave more emphasis to the document containing the passage. All passages were ranked according to their relevance, with the restriction that only three passages were allowed from the same document. The final result is top 1000 passages for a given query.

4 Results and Discussion

We conducted several runs during the experiment, taking different combinations of the techniques presented above. As shown in Figure 1, most of the runs outperformed the submitted baseline. Among the techniques we applied to improve over the low baseline runs, query expansion worked well, and the improvement was statistically significant (based on T-test). However, comparing to blind relevance feedback, which was our high baseline, the interactive query expansion approach performed slightly better, but the difference was not statistically significant.

Document re-ranking worked, but not as effective as query expansion. In addition, it seems that our approach of asking the user to rank their preference to the source of HARD sub-collections in general hurts performance (see the decrease of mean average precision in the run "interactive query expansion + strong source re-ranking" in Figure 1). We need further analysis to know the exact reason of this adverse effect. One possibility could be that the users actually did not know much about the documents in the sub-collection, so their preferences provided to us were not reliable in this case.

However, using the genre preference provided in meta data did help the retrieval performance a little, although there was no statistically significant difference between using and not using it, probably due to sparsity of genre preferences in the HARD topics. The potential advantage of genre preferences provided in meta data over the information about source preferences might have its explanation in that the genre information is easier for the user to determine, and recollect during the relevant assessment. Again, we need further study of the effect of using the genre information.

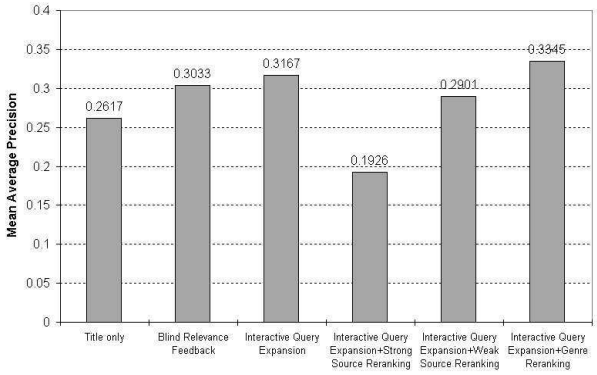

Figure 1: The comparison between different runs on document retrieval against two baseline document retrievals.

We performed failure analysis on the interactive query expansion run. Figure 2 shows the difference of average precisions between interactive query expansion and blind relevance feedback based on individual topics. The topics are arranged in the increasing order of the number of relevant documents in the

top 10 retrieved documents. In general, interactive query expansion performed better in the topics at the left end of the X axis, where these topics have none or very low number of relevant documents in the top 10 retrieved documents (e.g., topics 84 to 77). The improvement about 80% is statistically significant. The blind relevance feedback approach performed relatively better when the number of relevant documents in the top 10 retrieved documents increased, especially when most of the top 10 retrieved documents were relevant (i.e., topics 53 to 229).

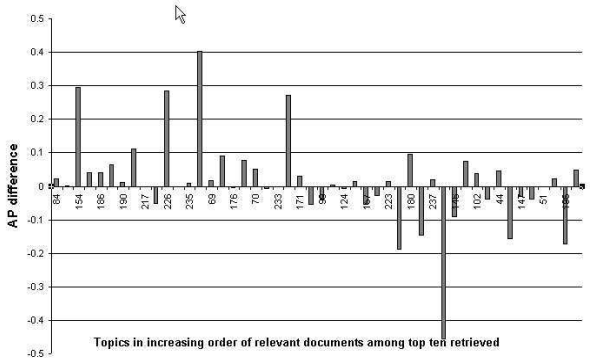

Figure 2: The comparison between interactive query expansion and blind relevance feedback on individual topic level. The upper half of the Y axis indicates that the interactive query expansion is better, whereas the lower half of the Y axis indicates that the blind relevance feedback is better.

It seems that it is helpful to have users' involvement when there is no or few relevant documents at the top of the ranked list, which is the exact place where blind relevance feedback could not perform well. In this case, the users' selection of clusters, even the non-relevant clusters, actually helps to remove some noise that could affect the retrieval effectiveness.

Therefore, a hybrid approach to query expansion is probably ideal. Interactions with users can be used for topics that have none or few relevant documents at the top, and blind relevance feedback should be used when most top ranked documents are relevant, which saves the trouble of having users make relevance judgments. The key here is whether or not we can somehow predict the number of relevant documents in the top ranked documents.

Our passage retrieval module performed reasonably well. In our passage retrieval run, 33 of 42 topics that required passages to be retrieved had the R-precisions above the median, and 6 topics had the best scores.

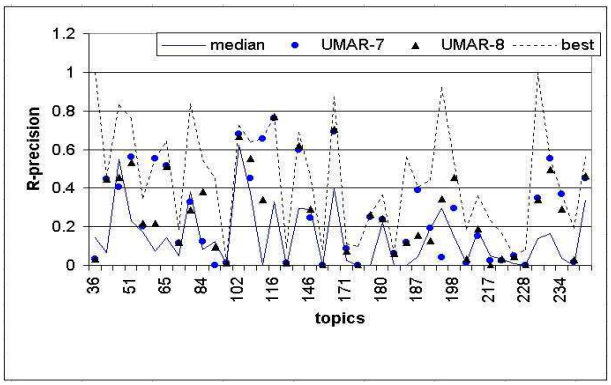

Figure 3: The performance of our passage retrieval runs among all submitted runs.

The run that tested ranked list merging was based on passage retrieval. UMAR-8 was the run that combined the ranked list of our passage retrieval (UMAR-7) and the passage retrieval version of our baseline run. Compared to the result of UMAR-7, UMAR-8 had two more topics with R-precision higher than the medians, but had much less topic among the best (only 1 topic for UMAR-8). One possible reason could be the suboptimal parameter setting used in the merging. Due to lack of training data, the parameters were chosen ad-hoc.

5 Conclusion

In this paper, we talked about our participation in the HARD track. Our approach in this year experiment was to explore the possibility of developing an automated HARD retrieval model by leveraging existing models and theories about information need negotiation in information science. Our initial analysis of the results indicates that this is a promising approach.

Although we have not finished analyzing our results yet, there are already some interesting lessons learned. The first one is about user involvement in the process. At least at the query expansion part, it is not the case that user's involvement would always improve retrieval effectiveness. When there are not many relevant documents at the top of the ranked list, asking user to perform cluster selection

is a good idea. Actually in this situation, there is no risk in asking for user's help. However, it is definitely not a good idea to ask the user select clusters if about half of the documents are relevant, and the other half is noise. Designing a mechanism to automatically identify when to ask the user, and when not to, will be one of the foci of our further exploration of the HARD retrieval model.

The second lesson learned is about the questions to the users. Our experience with the questions regarding users' preference to sub-collections indicates the importance of asking the right questions. The questions should be about the type of information that the users know, and also can easily express in their answers. If the questions actually force the users to make decisions that they do not fully understand, it probably is harmful to use the collected answers in the automated process.

Overall it was fun to participate in the HARD experiment, and we are looking forward to HARD 2004.

Acknowledgements

We thank Doug Oard and Eileen Abels for many long discussions about ideas, problems, and results; Zhou Liang and Ed Hovy at ISI University of Southern California for providing the GOSP Tool; James Allan for organizing the HARD track, and folks at NIST and LDC for providing topics and relevance judgments. This work has been supported in part by DARPA cooperative agreement N66001-00-2-8910

References

Eileen Abels. 1996. The Email Reference Interview. *RQ*, 35(3):345–358.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003a. Cross Language Headline Generation for Hindi. *This volume*.

Bonnie J. Dorr, Daqing He, Jun Luo, Douglas W. Oard, Richard Schwartz, Jianqiang Wang, and David Zajic. 2003b. iCLEF 2003 at Maryland: Translation Selection and Document Selection. In *Proceedings of CLEF'03*.

Daqing He, Jianqiang Wang, Douglas W. Oard, and Michael Nossal. 2002. Comparing user-assisted and Automatic Query Translation. In *Proceedings of CLEF'02*.

Marti A. Hearst and J. O. Pedersen. 1996. Reexamining the Cluster Hypothesis: Scatter/Gatherer on Retrieval Results. In *Proceedings of ACM SIGIR'96*, pages 76–84.

Jaap Kamps, Christof Monz, and Maarten de Rijke. 2002. Combining Evidence for Cross-Language Information Retrieval. In *Proceedings of CLEF'02*.

Anton Leuski and James Allan. 2000. Lighthouse: showing the way to relevant information. In Steven F. Roth and Daniel A. Keim, editors, *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)*, pages 125 – 130. IEEE Computer Society, October.

Gary Marchionini. 1995. *Information seeking in electronic environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger . In *Porceedings of Empirical Methods in Natural Language Processing Conference*.

R. S. Taylor. 1968. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29:178–94.

L. Zhou and E. Hovy. 2003. A Web-Trained Extraction Summarization System. In *Proceedings of the HLT-NAACL conference*, May.