ClickIR: Text Retrieval using a Dynamic Hypertext Interface

Richard C. Bodner and Mark H. Chignell

Interactive Media Laboratory
University of Toronto
5 King's College Road
Toronto, Ontario
Canada M5S 3G8
{rbodner, chignel}@mie.utoronto.ca

Abstract

In this report we describe our model of dynamic hypertext and how the ClickIR system uses this model to assist users in interactive search. The system was used in both the ad hoc task and the interactive track. In the context of the ad hoc task we were interested in the effects relevance feedback would have on our system. Comparison of ClickIR performance with and without relevance feedback showed that relevance feedback was critical in boosting the performance of the system from below median performance to the upper rank of TREC-7 systems. In the interactive track we compared the ClickIR (experimental) system where the tasks of querying and browsing were integrated, with a system which closely approximated a Web search engine, where the task of querying is separated from the task of browsing a list of hits. A trade-off between recall and precision was observed, with ClickIR leading to significantly greater recall, but at the expense of significantly lower precision and longer time taken to perform the task.

1.0 Introduction

The Interactive Media Lab at the University of Toronto participated in the manual ad hoc task and interactive track of TREC-7, following on from earlier participation in TREC-3 and TREC-4 (Charoenkitkarn *et al*., 1995; Charoenkitkarn *et al*., 1996). Our approach in these studies has been to support a person's decision-making ability while relieving some of the cognitive load placed on a person during searching (e.g., the tasks of querying and browsing). In TREC-3 and TREC-4 we used a query-based markup approach where searchers could mark up queries directly on the text of documents through click and drag operations. Since that time we have moved from using systems where queries are marked up on text to systems where queries are inferred from selections of text within documents. Instead of expressing queries, the searcher then simply has to click on sections of text to indicate his or her interest, and the system infers a query. For TREC-7, we were interested in comparing our model of dynamic hypertext with the functionality of a typical Web search engine. Our model of dynamic hypertext attempts to blend the tasks of querying and browsing whereas a standard search engine typically separates these tasks.

2.0 Dynamic Hypertext Information Retrieval Model

Standard, static hypertext documents have a number of well-known problems such as link maintenance, the lost-in-hyperspace problem (Conklin, 1987), and a finite structure. The last of these problems is probably the greatest of all since an author of a hypertext document cannot implement all possible links required by all possible visitors to that document. This problem leads to many unconnected small groups of hypertext documents. These groups are difficult to navigate. This has lead to the need for search engines to assist in navigation. The use of search engines creates a "spiky" navigation pattern (see Campagnini and Ehrlich, 1989; and Parunak, 1989 for a description). Due to this spiky pattern, users navigate in two modes:

searching (e.g., querying a search engine) and browsing (e.g., reviewing the documents retrieved based by the previous query).

Our model of dynamic hypertext attempts to blend these two modes together via the user interface. In our approach there is no notion of a static link; links are created on the fly based on knowledge of the corpus and the interests of the user. We infer the user's interests by recording his/her interaction (links clicked on) with the system. Various query formulation algorithms can be used to infer a query once a link or a section of text has been selected. In the current implementation of the ClickIR system, the sentence that the link occurs in is sent to the search engine as a query. It is assumed in this approach that the user selected the link due to an interest in the content surrounding the link. Thus the dynamic hypertext acts like a form of sentence-based relevance feedback.

In ClickIR, markup of links within text is dynamic. Words are selected as links based on the user's previous interactions with the system. The words in the previous queries are used as seeds for selecting the future links. A "tail" representing a weighted average of the three most recent queries is used in selecting terms to be highlighted as links, and in modifying the query formed after a link is selected. This "tail" is so named because it adds inertia to the process of switching from one topic or class of query to another. The number of previous queries used in forming the tail is a parameter that can be changed. The weighting vector used to capture the diminishing importance of older queries relative to more recent queries can also be changed. The value currently used to weight the important of past queries (thereby defining the "tail size") was arrived at through informal experiments carried out in our lab. For a further description of our dynamic hypertext model see Bodner *et al*., 1997 and Tam *et al*., 1997.

3.0 System Descriptions

The Interactive Media Lab implemented two systems specifically for participation in TREC-7. For the ad hoc task, only the experimental system (ClickIR) was used. In contrast, the interactive track required both a control and an experimental system. Both the control and experimental systems used the Inquery (version 3.1) search engine software (Caltan *et al*., 1992). Our systems provide a hypertext user interface to the search engine. This was accomplished through the use of CGI scripts. The scripts convert the user's interactions with the system into queries, which are sent to Inquery. The query results are then converted into hypertext documents. Neither the experimental nor the control system allowed the use of any Boolean or special search operator. Users were only allowed to enter natural language queries.

3.1 The Control System

The control system used in the interactive track was designed to mimic the hypertext interface provided by most search engines on the Web today. The typical search engine interface separates the tasks of querying and browsing. The user must constantly switch between modes during a search session. In our control system the user was presented with a startup screen where he/she could enter an initial query. From the initial query a resulting list of document titles, which were linked to the full document text, were presented in ascending rank order (provided by Inquery). On this screen, the user could also enter new queries to continue the search session (see Figure 1). When the user selected a title link, he/she entered the document view screen. On this screen there was a link that users could click on to mark the current document as being relevant to the current search and there was another link which led to a review of the user's relevant document list. The user iterated between these two screens during the search session (task switching between querying and browsing).

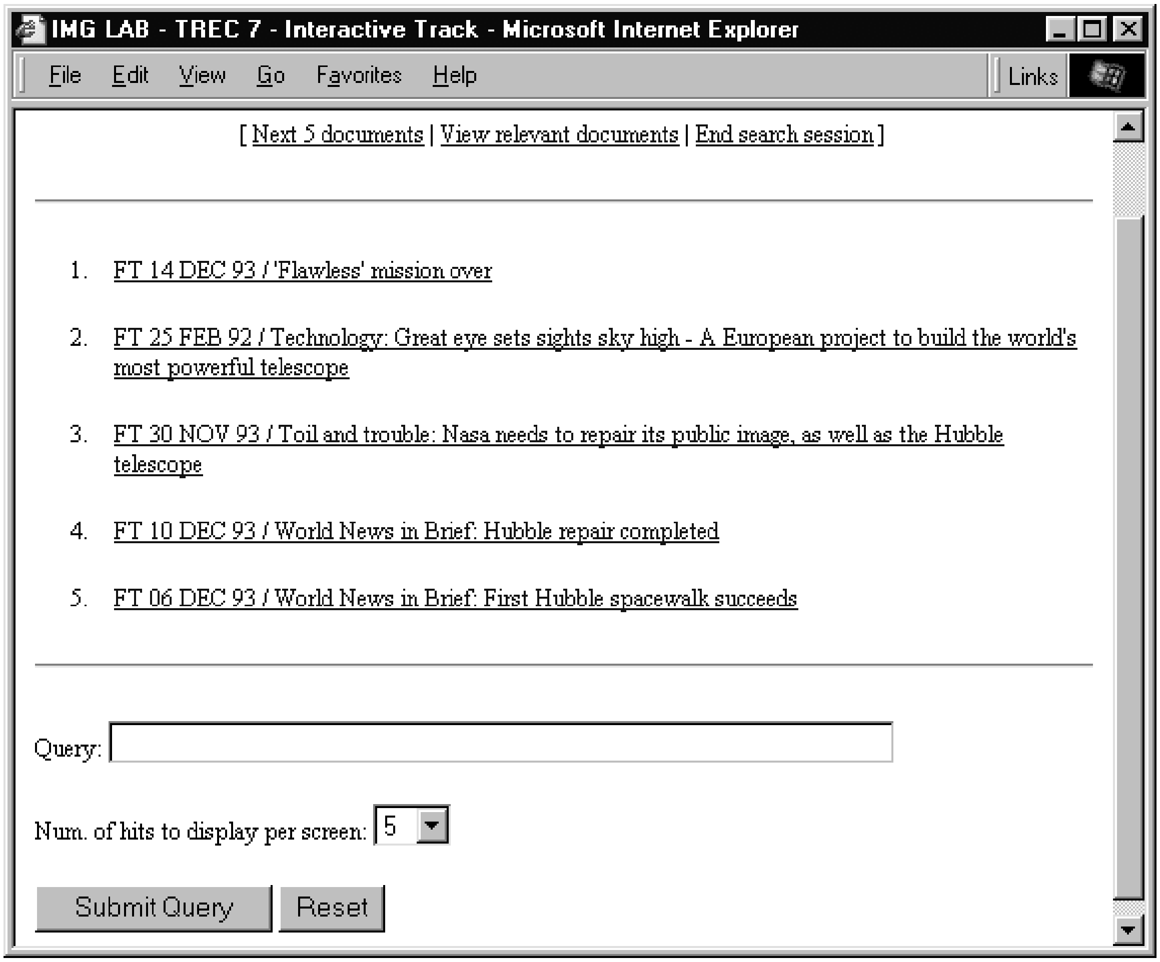

FIGURE 1. Result list screen for the control system.

3.2 The Experimental System

The experimental system, known as ClickIR, was used in both the interactive track and the manual ad hoc task. The experimental system implemented our model of dynamic hypertext (described above in Section 2.0) in which the tasks of querying and browsing are blended together. As with the control system, the user was presented with an initial query screen. The query results screen (see Figure 2) displayed the entire text of one or more documents. Although the user could select the number of documents to display per results page, the default was to display two documents. Through informal pilot studies this number of screens was found to be easiest to handle, both in terms of scrolling and in terms of minimizing information overload. The CGI scripts collected the results from Inquery, and used the document context and terms found in the user's queries as sources of information to guide the markup of the documents, which were then presented as dynamic hypertext documents. The user queried the system by clicking on hypertext links in the documents. The system also provided a form to enter a new query or expand the current query. This form was included because the conditions imposed by TREC did not allow for the system to be primed with terms in order to provide more appropriate links to the user. As with the control system the user could "page" through the query result set, mark documents as being relevant, and review their relevant document list. The user did not switch between querying and browsing tasks in the experimental system.

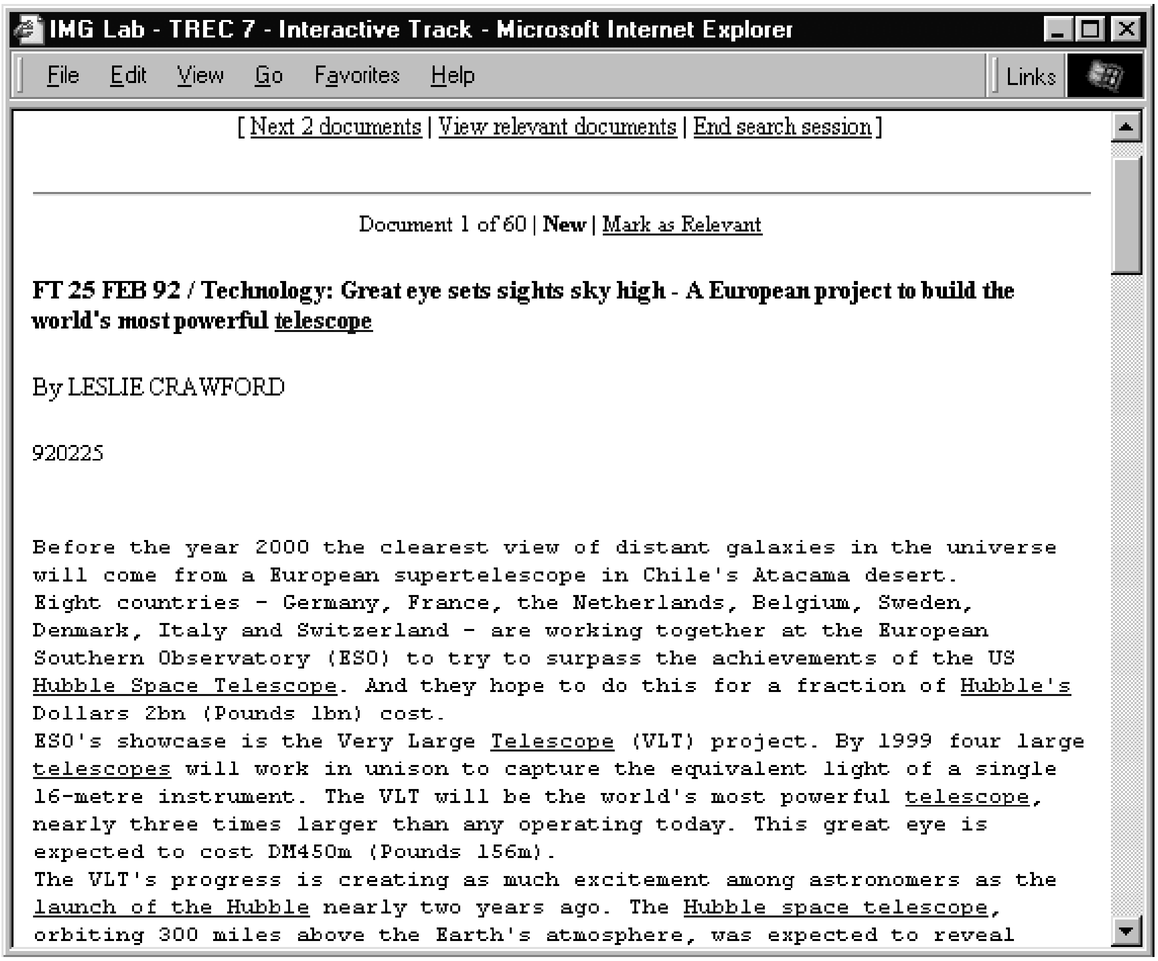

FIGURE 2. Document view screen for the experimental system (ClickIR).

One of the issues with our experimental system that we investigated at TREC-7 was phrase generation. In previous versions of our system, links were simply single terms (words) and terms were selected as links based on the product of term frequency and inverse document frequency (TF*IDF). Using this metric an upper and lower threshold could be set in order to limit or expand the number of links that were generated. This threshold method did not adequately address the problem of multiple occurrences of a term being highlighted multiple times in the same document. A user with only standard hypertext experience would then have difficulty distinguishing between the different occurrences of the term (i.e., the term/link would appear to point to the same end node, given experience with how standard hypertext tends to work). To try to solve this problem a phrase generation technique was used. We expected that the use of a phrase instead of individual terms would help users by providing more discriminable links. For example, if the term "retrieval" appeared in multiple locations within a document, a user might assume that each occurrence points to the same endpoint. In contrast, if one occurrence of the term occurred in the phrase "information retrieval" and another occurred in "retrieval of documents", the user should then be able to distinguish the different links.

Given time contrasts, we could only implement a phrase generation method based on simple heuristics to create the phrases. The heuristics operated on the following assumptions:

- A phrase is two or more words containing two or more content bearing terms. A content bearing term is defined simply as a term that is not a stopword.

- A phrase can have at most two stopwords between content bearing terms.
- A phrase must begin and end on content bearing terms.

4.0 Manual Ad Hoc Task

Although ClickIR was not designed for handling tasks such as TREC manual ad hoc, we used ClickIR in that task in an attempt to understand how relevance feedback would affect the performance of our system. The two runs we submitted to TREC were `uoftimgr` and `uoftimgu`. Only the first of these runs used relevance feedback.

As described in Section 2.0, our dynamic hypertext information retrieval model uses sentence-based relevance feedback interactively during search. The interactive search was combined with a batch search in order to conform with the requirements of the TREC manual ad hoc task, as will be further discussed in Section 4.1 of this paper.

4.1 Query Generation Process

Since ClickIR is an interactive system, the searcher was given approximately 15 minutes per topic to browse the document collection. As the searcher browsed the collection, the links that he/she clicked on were recorded (called the "interaction record"). The searcher was also asked to select documents that "seemed" relevant to the topic. Given the amount of data in the TREC-7 ad hoc collection, we felt that searchers would be able to get a sense of the type of documents the collection contained even if they did not find a large number of relevant documents.

The searcher's interaction record was then used to build a query by collecting all the sentences for the links that were clicked on (selected). The resulting query was then sent to Inquery and the top 1000 documents were selected. This was how the uoftimgu queries were generated. The uoftimgr queries also contained the queries generated from the interaction record, but in addition, the documents marked as relevant during the search were used for relevance feedback.

4.2 Results and Discussion

We found a significant difference in average precision between the uoftimgr and uoftimgu runs ($t[49]=2.18$, $p<.05$). As expected, the relevance feedback runs performed better than the simple interaction record runs. The average precision for all topics for uoftimgr was 0.276 and for uoftimgu the average precision was 0.245. Overall, the uoftimgr run retrieved 60.3% of the total relevant documents identified by TREC and 5.6% of the total documents retrieved by the system were relevant. The uoftimgr run retrieved 55.3% of the relevant documents and 5.2% of the total documents retrieved were relevant. Figure 3 shows the recall-precision curves for both runs. While this difference may not sound like much, in the context of the manual ad hoc task in TREC-7 it meant the difference between a system that was average and a system that was one of the best.

An alternative way of viewing the effect of relevance feedback is shown in Table 1. When relevance feedback was used, the average precision was above the median (i.e., the median average precision score for all 17 research groups) for 31 or 62% of the 50 topics. However, when relevance feedback was not used, the system performed better than the median only 44% of the time (or for 22 of the 50 topics). Further evidence of the major impact of relevance feedback was shown by the fact that, on average, five additional relevant documents were retrieved using relevance feedback (an average of 56.9 with, versus 51.7 without, relevance feedback) which was a statistically significant difference ($t[49]=2.04$, $p<.05$).

TABLE 1. System performance based on average precision median for all topics in the manual ad hoc task.

		Without Relevance Feedback (uoftimgu)		
		Better	**Worse**	
With Relevance Feedback (uoftimgr)	**Better**	18	13	31/50
	Worse	4	14	18/50
		22/50	27/50	

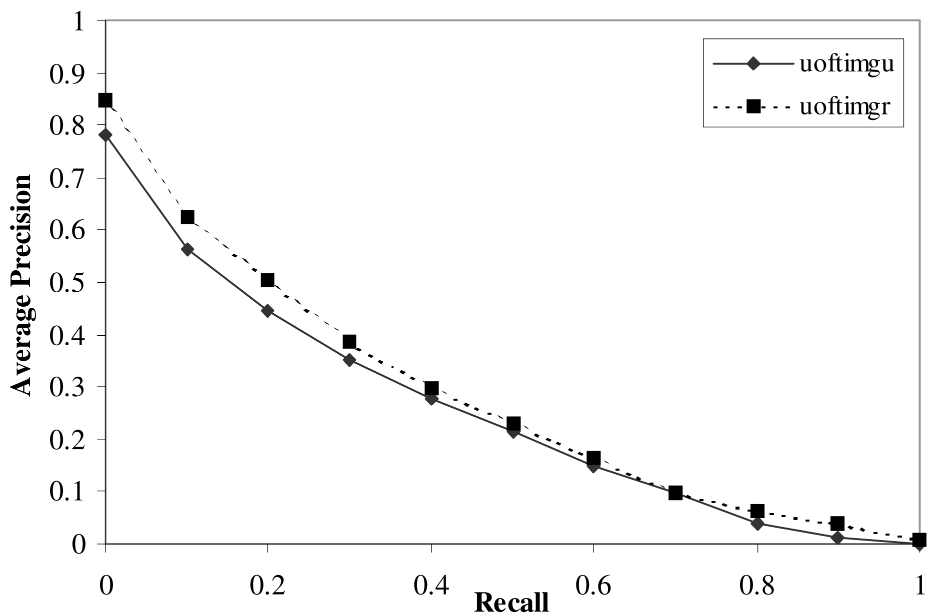

FIGURE 3. Recall-precision curves for uoftimgu and uoftimgr runs.

5.0 Interactive Track

For our participation in the interactive track, we were interested in comparing the dynamic hypertext information retrieval model implemented as ClickIR (the experimental system) with an interface which separated the tasks of querying and browsing such as found in current Web search engines (the control system).

5.1 Experimental Design

The design of the experiment in this section of the study followed the design required of the TREC-7 interactive track participants.

The eight search topics were ordered into two distinct sequences of four topics each. Within each sequence, the ordering of the topics was fixed, but the ordering of the sequences was counterbalanced, so that half of the subjects worked on one sequence of four topics followed by the other, while the remaining subjects worked on the two sequences in the reverse order.

The other factor that was manipulated in the experiment was the type of system used. The interactive experiment used ClickIR as the experimental system (as described in section 3.2). Note that ClickIR is referred to as "System A" in our online reports and the control system is referred to as "System B". The performance with this experimental system was contrasted with the control system (described earlier in Section 3.1). Half of the subjects used the experimental condition first for four trials (one of the two sequences) followed by the control condition (with the four tasks from the other sequence) and half used the two systems in the reverse order.

There were a total of four possible combinations of sequence and system. Each subject was presented with two of these four possible combinations for a total of eight trials each (two combinations with four search topics per combination).

For the purposes of analysis and interpretation, the search topics were classified according to their average level of difficulty as found by averaging the results of all the interactive track participants. Topics with an average recall of 0.3 or better were classified as "easy" and topics with lower recall were classified as "difficult". The analysis then focused on the impact of system and topic difficulty on search results.

Measures collected during the search included the number of documents viewed, the overall time taken, the time taken to find the first relevant document, and the recall and precision achieved. In addition, a standard set of six questions (involving ratings on a five point rating scale) was completed at the end of each search.

In addition to the measures collected during and immediately after the searches, a pre-experiment and a post-experiment questionnaire was administered. The pre-experiment measurements included the word associations (FA-1) task that was selected by the organizers of the track.

5.2 Subjects

The eight subjects that participated in the experiment were selected from the University of Toronto undergraduate and graduate student populations. There were three female subjects with an average age of 24 and an average FA-1 pooled score of 38. The five male subjects had an average age of 32 and an average (mean) FA-1 pooled score of 32.8. None of the subjects had previously participated in a TREC searching study and all of the subjects reported having a high degree of experience searching the Web.

5.3 Results

Multivariate analysis of variance was used to assess overall effects. There was no significant interaction between system and topic difficulty. However, there were significant main effects for topic difficulty and for system. The nature of these effects was then assessed using a fully within two factor (system by topic difficulty) analysis of variance for each of the measures. For topic difficulty, there were significant differences on a number of the measures which were in the expected direction, indicating that our results agreed with the general ordering of topic difficulty that was found across the groups participating in the TREC-7 interactive track.

For the system main effect, there were significant ($p<.05$) effects for recall, precision, and total search time and borderline significant ($p<.10$) effects for time to first relevant document, and for questions 2 through 6 of the post-search questionnaire. The experimental system had significantly higher ($F[1,7]=42.74$, $p<.001$), recall than the control (mean of .41 vs. a mean of .37), but at the expense of significantly lower ($F[1,7]=35.66$, $p<.001$) instance precision (mean of .65 for instance precision for the experimental system versus .70 for the control system). There was no significant difference in overall time taken by the two systems (with searches frequently taking the complete 15 minute maximum that was allotted), neither was there a significant difference in the time taken to find the first relevant document ($F<1$ in both cases).

ANOVA tests on the post-search questions should be interpreted cautiously, since the 5-point Likert scale responses on each question are not continuous normally distributed variates. However, the ANOVA results are presented here in the spirit of generating hypotheses for further research (with a skeptical stance concerning the precise values of F and p for each test being assumed). That being said, the effect of system on question 2 (“was it easy to get started on this search”) was borderline significant ($p<.10$) with a tendency for subjects to judge the control system as being easier to get started with. There was no significant difference in satisfaction with the results obtained when using the two systems (F was approximately 1 for the comparison on question 4). There was also no significant difference in their confidence that they had identified all the different instances of each topic (question 5). There was however a borderline ($p<.10$) effect for question 6 (“did you have enough time to do an effective search”) with more participants tending to feel that they had enough time with the control system. From a methodological standpoint, it is interesting to note that this subjective question was more effective to a possible difference in how much time was needed to search using each system than was the actual measure itself.

From the exit questionnaires, subjects preferred the control system in terms of ease of learning (5/8). Ease of use was evenly split between the two systems, and 6 out of the 8 subject reported that they liked the experimental system the best.

5.4 Discussion

These results indicate that the experimental system promoted recall at the expense of precision. Earlier studies in our laboratory have found that search experts tend to have higher precision, but lower recall, than experts (Charoenkitkarn, 1996; Golovchinsky, 1997). In contrast, this studies showed (using a within-subjects design) that two different interfaces tended to move participants (as a group) to different points on a trade-off between recall and precision. Golovchinsky (1997, p. 120) classified his experimental subjects as “skimmers” (i.e., people who make many interactions with the system during a session) or “readers” (i.e., people who spend more time reading articles and making careful judgements). One intriguing possibility for future research is that the point and click nature of the experimental system encouraged a high degree of interactivity or “skimming” and that as with Golovchinsky's subjects, higher recall resulted.

An alternative viewpoint that deserves further study stems from the fact that the experimental system made more text available for a longer time within the experiment. As a result, there may be more incidental learning about the topics in the text in the dynamic hypertext version of the retrieval system than there is with a standard Web search interface.

The user interface for the experimental system was designed for ease of use. Its browsing interface can function as both a dynamic hypertext system and as a user interface for text retrieval. In an earlier study, the query-based dynamic hypertext was compared with a static hypertext (Tam, 1997; Tam *et al*., 1997). Tam found that searching for information using the experimental interface was easier for computer/domain novices than searching for the information in a static hypertext version of the information. For experts, however, there was no significant difference in the level of performance obtained with the dynamic version of the hypertext (experts were neither helped nor hindered by the interface).

In the present study, the control system represented a highly familiar search engine interface for all of the experimental subjects. Thus it might be supposed that the novel experimental system would be at a natural disadvantage when compared with the familiar control system. In order to discount this possibility, a longitudinal study would be required where subjects use the experimental system for an extended period of time to see if their performance improves as they gain more experience with the system. However, such an analysis was outside the scope of the experiment defined for participants in the interactive track of TREC-7.

The results obtained in this study must be considered tentative due to the relatively small sample of subjects used and to the small amount of experience they had with the experimental system in contrast to their experience with the search engine interface. In spite of this caveat, there is a clear tendency for the experimental interface to promote recall at the expense of precision. This suggests that the experimental interface will prove useful in situations where recall is emphasized. Possible areas where this is the case may include patent searches and exhaustive literature reviews. In addition, it is expected that people will benefit from reading more document text, as is likely to occur in the dynamic hypertext interface. However, the demonstration of learning as a supplement to retrieval was outside the scope of this study and represents a hypothesis to be tested in subsequent research.

6.0 Conclusions

This study was the first time that a query-based dynamic hypertext interface has been tested under TREC conditions. Earlier participation by our research group had used visual mark-up based querying, which stopped short of the point and click link selection method introduced in this study of large scale text retrieval. The tendency for different interfaces/systems to produce a trade-off in recall vs. precision may provide a useful stimulus for further research. Understanding how and why this trade-off occurs may provide fundamental insights into how search behaviour changes depending on the type of system and interface. One suggestion is that the increased emphasis on recall at the expense of precision, with the dynamic hypertext system, is due to the increased availability of the text, and to the way in which query intent is expressed with respect to the text.

The experimental system yielded surprisingly good results in the ad hoc section of the TREC-7 competition, which is a notable result for a system that emphasizes interactive search over complex computational techniques. However, it should be noted that two sets of outputs were submitted to TREC, those that were output by the experimental system, and those that were "boosted" through relevance feedback. That boosting was based on the relevance judgements (document selections) made by the two subjects used in the ad hoc phase of the study. The comparison of the regular and boosted results showed that boosting with relevance feedback did in fact significantly improve the average precision, and is a useful supplement for a query-based dynamic hypertext system, where relevance judgements can be collected as an unobtrusive by-product of the interaction with the system.

The present findings concerning relevance feedback demonstrate that it makes an important improvement to the effectiveness of query-based dynamic hypertext. In addition, the findings support the notion that differences between the average precision obtained by different research groups are small enough that small differences in average precision can lead to fairly major changes in the ranking obtained in the TREC conference.

Acknowledgements

The authors would like to thanks the subjects that took part in the Interactive track portion of the study. We would also like to thank Rick Kopak for his assistance in conducting the manual ad hoc searches. This project was funded in part by National Science and Research Council of Canada (NSERC) and Communications and Information Technology of Ontario (CITO).

References

Bodner, R., Chignell, M. and Tam, J. (1997) Website authoring using dynamic hypertext. In *Proceedings of Webnet'97*, Toronto: Association for the Advancement of Computing in Education, 59-64.

Caltan, J. P., Croft, W. B., and Harding, S. M. (1992). The INQUERY retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, 78-82.

Campagnini, F. R. and Ehrlich, K. (1989) Information retrieval using a hypertext-based help system. *ACM Transactions on Office Information Systems*, 7(3), 271-291.

Charoenkitkarn, N. (1996). *The effect of markup-querying on search pattern and performance in large-scale text retrieval*. Unpublished Ph.D. dissertation. Department of Industrial Engineering, University of Toronto.

Charoenkitkarn, N., Chignell, M.H., and Golovchinsky, G. (1995). Interactive exploration as a formal text retrieval method: How well can interactivity compensate for unsophisticated retrieval algorithms?

Pages 179-199. D.K. Harman (Ed.), In *Proceedings of the Third Text Retrieval Conference (TREC-3)*. National Institute of Standards Special Publication 500-225. Gaithersburg, Maryland.

Charoenkitkarn, N., Chignell, M.H., and Golovchinsky, G. (1996). Is recall relevant? An analysis of how user interface conditions affect strategies and performance in large scale text retrieval. In D.K. Harman (Ed.), *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*. National Institute of Standards, Gaithersburg, Maryland.

Conklin, J. (1987, September) Hypertext: an introduction and survey. *Computer*, 20(9), 17-41.

Golovchinsky, G. (1997). *From information retrieval to hypertext and back again: the role of interaction in the information exploration interface*. Ph.D. Dissertation, Department of Mechanical and Industrial Engineering, University of Toronto.

Parunak, H. V. D. (1989) Hypermedia topologies and user navigation. In *Proceedings of Hypertext'89*. Pittsburgh, PA: ACM Press, 43-50.

Tam, J. (1997). *Design and evaluation of web-based dynamic hypertext*. Ph.D. Dissertation, Department of Mechanical and Industrial Engineering, University of Toronto.

Tam, J., Bodner, R. and Chignell, M. (1997) Dynamic hypertext benefits novices in question answering. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, 350-354.