Evaluating a Visual Information Retrieval Interface: AspInquery at TREC-6

Russell Swan James Allan Don Byrd
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003
{swan, allan, dbyrd}@cs.umass.edu

ABSTRACT

We built two Information Retrieval systems that were targeted for the TREC-6 "aspect oriented" retrieval track. The systems were built to test the usefulness of different visualizations in an interactive IR setting —in particular, an "aspect window" for the chosen task, and a 3-D visualization of document inter-relationships. We studied 24 users of the system and report the usage of different user interface elements, whether experienced users outperformed novices, and whether spatial reasoning ability was a good predictor of effective use of 3-D.

INTRODUCTION

Much if not most usage of Information Retrieval (IR) is in an interactive setting, where an individual must use an IR system to obtain information that they need. But historically, most IR evaluations are based on comparing batch mode runs of IR systems on fixed corpora and measuring precision (the percentage of documents that are retrieved that are relevant) and recall (the percentage of relevant documents that are retrieved). IR systems were expensive and used primarily by trained professionals for a very specific task. With the explosion of information available, information retrieval, exploration and organization are becoming tasks that users of very diverse backgrounds are required to perform.

In order to help users with these tasks, we are interested in using visualizations to create targeted systems, where systems can be targeted to either a specific task, or to a specific type of user for that task. By using interactive graphical displays we hope to be able to convey more information to users at a higher rate about document content, relevance and relatedness by the use of position and color than can be given by lists and numeric scores. We feel that it is possible to create specialized interfaces to systems in order to increase their suitability for a task over a non specialized system. We also feel that we can design and build systems that will increase the effectiveness for specified types of users. Some questions we hoped to answer were: how much better are skilled searchers (librarians) compared with novice searchers? Will the two groups react differently to different visualizations, i.e., will a certain visualization increase the effectiveness of skilled searchers more than novice searchers, or will a different visualization be of no use to skilled searchers but aid novice searchers?

In this study, we investigated exactly those questions. The work was driven by the TREC-6 (Text REtrieval Conference) Interactive Track, an evaluation of "aspect oriented information retrieval," wherein users are tasked with identifying as many "aspects" of relevance to a query as they can. For example, in a query about ferry sinkings in the news, the task was to find a list of all ferries that sank, not to find all documents about ferry sinkings. In order to perform this task well it was important to not just find the information, but organize it as it was found, determining if a new document represented a new aspect of the topic, or if it had already been covered. The structure of our experiments was determined to a large extent by the TREC-6 guidelines; they are explained in more detail below.

SYSTEM

We used three systems for the experiments discussed in this study:

1. ZPRISE (ZP) is a basic GUI information retrieval system. This is the "control" system for our experiments.
2. AspInquery (AI) is a GUI implementation of Inquery that includes an "aspect window" to help with the task. The core of AspInquery is a basic GUI similar to ZPRISE.
3. AspInquery Plus (AI+) is an extension of (2) that includes a 3-D visualization of document relations.

The baseline system for our experiments was ZPRISE, supplied by the National Institute of Standards and Technology (NIST). ZPRISE uses a straightforward user interface much like that used by most Internet search engines: it has an area for typing in a query, a window for displaying a ranked list of documents, and a window for viewing a document of interest. When the full text of a document is viewed, query terms contained in the document are highlighted. Documents were marked as relevant either by clicking on a button in the ranked list, or clicking on a

button in the document window.

Experimental Systems

Our system consisted of the Inquery search engine[1] with a new interface. Our basic user interface has much in common with the ZPRISE interface, differing in two significant ways: ZPRISE displays the query terms contained in a document after the headline but our system does not, and our system color codes whether a document has been viewed but ZPRISE does not. Specifically we write the headline information of a document in blue if it has not been viewed before, and purple if it has been seen. (This scheme was modeled after the default color scheme Web browsers use to show if a hypertext link has been followed or not.)

particular aspect can be stored. To help label the information, statistical analysis of word and phrase occurrences is used to decide what terms and phrases are most distinctive about a document or set of documents in an aspect. We provided an area for the user to manually assign additional keywords or labels if needed.

Each area of the aspect window has a colored border, a text field at the top for entering a descriptive label, and an automatically generated list of the five noun phrases that most distinguish the group of documents assigned to this aspect from the remainder of the collection. Figure 1 shows an example of the aspect window. The system shows two third area waiting for the next aspect. The first aspect contains

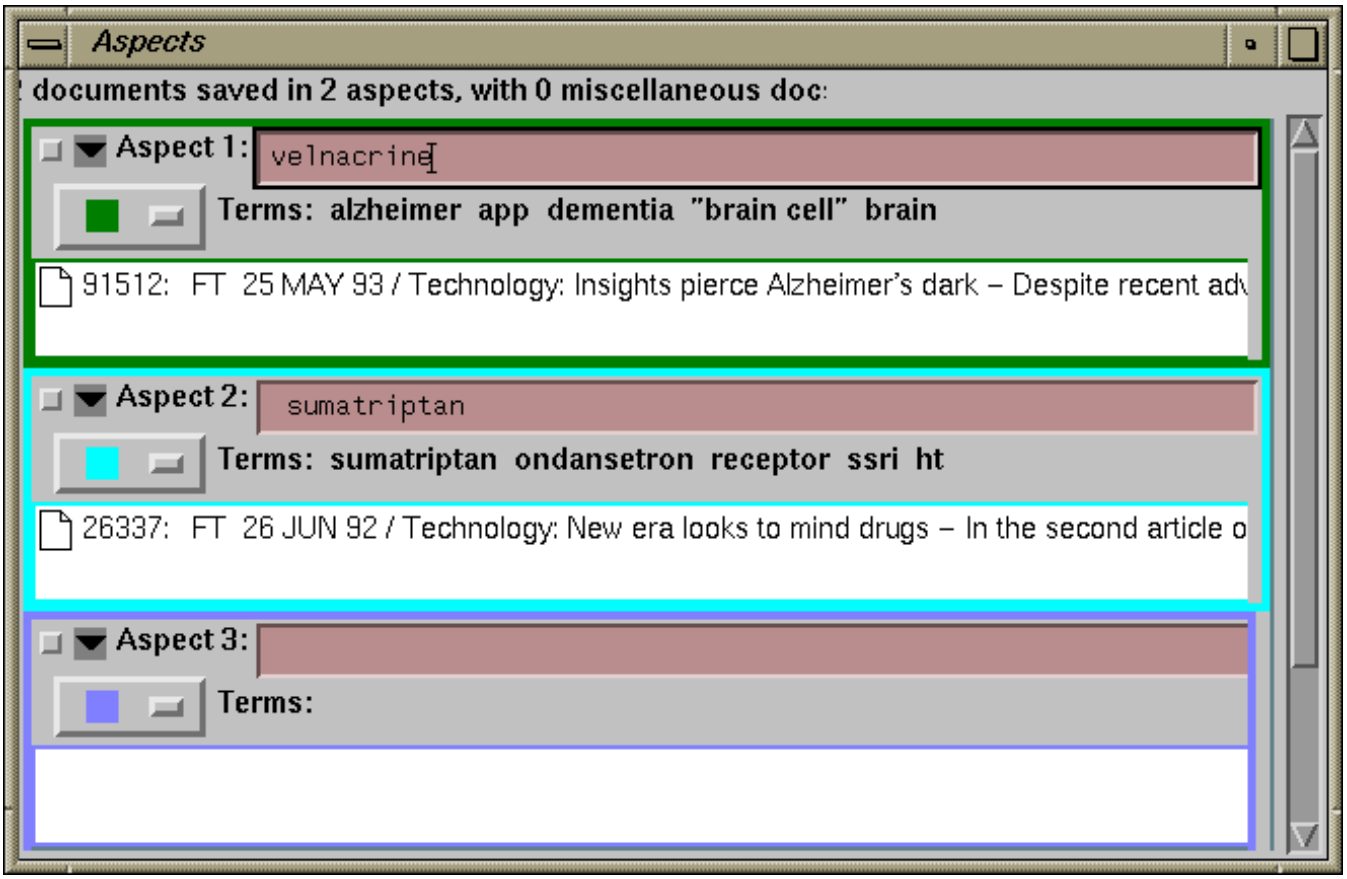

Figure 1. The Aspect Window

Aspect Window

For the aspect oriented retrieval task, the user must not only save documents relevant to the query, but must also keep track of which aspects have been already identified. With a standard IR system the user would need to mentally keep track of this information, or write it down on a scratch piece of paper or in an auxiliary text window. We implemented an "aspect window" tool to help with this task. The idea is to provide an area where documents on a one document, that the user entered into the aspect by dragging from the ranked list display into the aspect's document list. The system then analyzed the selected document and found five phrases that describe the aspect; the analyst manually added "velnacrine".

Visualization: AspInquery Plus

Another important step in the aspect oriented retrieval task is deciding (repeatedly) which document to look at next. In a ranked retrieval system the documents are presented in

the order of probability of relevance, so the user is more likely to encounter relevant documents at the top of the list than further down. The headline is generally used to decide if the full text is worth reviewing or not. Some systems[4,8], ZPRISE among them, give information about the query terms that appear in the document, expecting that they can be used to help decide whether to investigate further. But for an aspect retrieval task, the deciding point of whether to investigate a document further is not the information content, but the marginal information content—i.e., the information content in the context of what has already been seen. The Cluster Hypothesis[7] states that relevant documents tend to cluster, and it has been shown to be valid in top-ranked documents[2,5]. Aspects represent different forms of relevance, and we believe that they will group together within thegroups of documents (two aspects) already identified and a set of relevant documents.

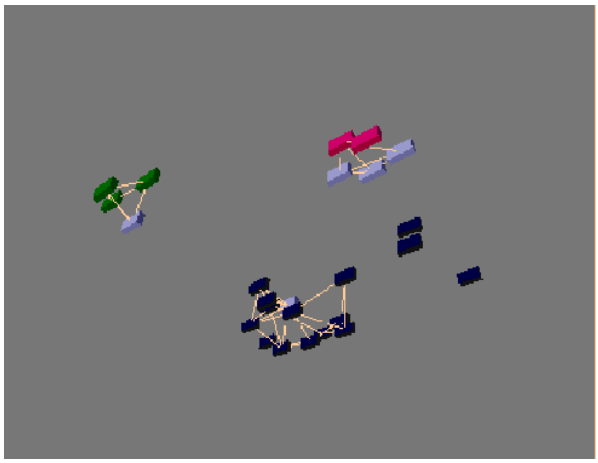

Figure 2. The 3-D Window

AspInquery Plus compares documents in an extremely high-dimensional space (approximately 400,000 for this collection) where each dimension corresponds to a feature in the collection and the distance is measured by the sine of the angle between the vectors. That space is collapsed to 3 dimensions for visualization using a spring embedding algorithm[6].

Documents that are nearby in 3-space are generally nearby in the high dimensional space also, meaning that they share information content to a considerable degree. For that reason, the 3-D display provides the user with information about whether the document is worth investigating further, helping the user to sort through documents more quickly. Documents in the 3-D window are persistent between queries: when new documents are retrieved they are colored light blue (light purple when read) and are placed in the 3-D window by the forces exerted from already placed documents. Figure 2 shows five newly retrieved documents in light gray. It is easy to see that three of these documents fall into a group of two previously seen documents (upper right of figure) and the other new documents fall into the small group in the upper left and the large group. A user who is under time pressure could use the 3-D display to decide that the unjudged document near that aspect is probably on the same aspect and so not worth examining. A retrieved document that is far from any already-marked aspect is more likely to be useful.

The three windows—result list, aspect, and 3-D—are tightly integrated. If a document is selected by a mouse click in any of the three windows, that document is highlighted in all windows in which it is visible. A document can be opened for viewing by double clicking in any of the three windows. The colors were coordinated between the windows: if a document has been saved to an aspect, that aspect's color is assigned to the document in the 3-D window and also displayed before the document in the list.

EXPERIMENT

Participants

We had a total of 24 participants in our user study. We were interested in how librarians perform search tasks as compared to a more general user population, so we divided our population equally between librarians and general users. Twelve university librarians (all with MLS degrees) were recruited for the study and four were placed in each experimental group. The general population was recruited by flyers distributed on campus. This group was primarily students (10 of 12 participants). The librarians were older, more highly educated, and had more years of searching experience.

Procedure

The basic unit for our experimental design was a block, each block having four users. Each user ran six topics, three with the experimental system and three with the control system. Two of the four users did the first three searches with the experimental system, and the other two users did the first three searches with the control system. Topic order was held constant. This Latin square design allows blocking on both topics and users, and the average of the diagonals gives an estimate of system-specific differences. This block design, the topics run, the order of the topics, and the use of ZPRISE as a control were all specified by NIST as part of the TREC Interactive Track.

We ran three groups each composed of two blocks, one block of general users and one block of librarians. This design allowed us to block on experienced/novice users in our assessment of the systems. The first group compared AI with ZP, the second group compared AI+ with ZP, and the third group compared AI with AI+ directly.

Before the searches, each participant filled out a questionnaire

to determine age, education, gender and computer experience, and took two psychometric tests[3], a test of verbal fluency (Controlled Associations, test FA-1) and a test for structural visualization (Paper Folding, test VZ-2).

Data Set and Measures

The corpus used was newspaper articles from the Financial Times, 1991-1994, approximately 200,000 articles total, a subset of the TREC collection. The measures used to evaluate performance of the different systems were aspect oriented precision and recall. Precision and recall are standard measures used for evaluating IR systems. The aspect oriented versions of these differ from the standard in that aspect oriented recall is the percentage of all aspects identified from the corpus that are contained in the saved set, and aspect oriented precision is the percentage of saved documents that contain information on one or more aspects.

RESULTS

Usage of Visualizations

Users were required to drag documents to the aspect window in order to save them. The aspect window was intended to help the user in categorizing and keeping track of what information had been seen before. All users were offered a piece of scratch paper when using ZPRISE, so in order to keep the protocol the same we also offered them a piece of scratch paper when using the two experimental systems. For the users in the groups that used ZP as a control, there were 48 searches with ZP, and 48 searches where an aspect window was used. Scratch sheets were used on 22 of the 48 searches with ZP, and 14 of the 48 searches on the experimental system, for a χ^2 value of 2.844 (NS). Librarians tended to use scratch sheets more than students (26 of 48 searches for librarians, 10 of 48 for students, $\chi^2 = 11.378$, $p < 0.005$). One subject always used a scratch sheet, and seven subjects never used one. For the eight participants that sometimes used them, scratch sheets were used on 11 of 24 AI(+) searches, and 19 of 24 ZP searches, for a χ^2 value of 5.689, $p < 0.025$.

The aspect window provided an area for labelling aspects, and suggested the top 5 terms and phrases that distinguished those aspects. Most aspects were labelled, but the suggested terms and phrases were used slightly less than half the time as labels.

The 3-D visualization was intended to provide an alternative to the ranked list for navigating the retrieved set of documents. Full text of documents could be retrieved by double clicking on the document title in the ranked list, or by double clicking on the document icon in the 3-D view. It was possible to perform the task without ever using the 3-D window, and from previous experience we knew usage would vary widely among participants, so we instrumented the 3-D window to record mouse clicks as a measure of usage. We hypothesized that usage of the 3-D window would correlate highly with test scores on Structural Visualization (test VZ-2). Figure 3 shows a scatterplot of VZ-2 score versus usage of the 3-D window.

The data fall into 3 clusters—a cluster labeled “A” that had a moderately high score on VZ-2 and used the window very little, a second group “B” that scored very highly in VZ-2 and used the window extensively, and a third group “C” that scored below average on VZ-2 but used the window extensively. Clusters “A” and “B” in isolation would be confirmation for our hypothesis, but cluster “C” is not what we expected. A possible explanation is that the individuals in cluster “A” have a natural ability with 3-D but limited experience with 3-D on computers, with mouse based interfaces, and with GUIs. The participants in cluster “C” on the other hand might be very comfortable with GUIs, mice, and 3-D interfaces. To test this we examined the scores of the participants on our entry questionnaire. We found that the users in Cluster “A” reported less experience with mouse based interfaces than the users in clusters “B” or “C” ($p < 0.05$), implying that whether or not a person uses a 3-D interface depends more on their familiarity with GUIs than with natural 3-D ability.

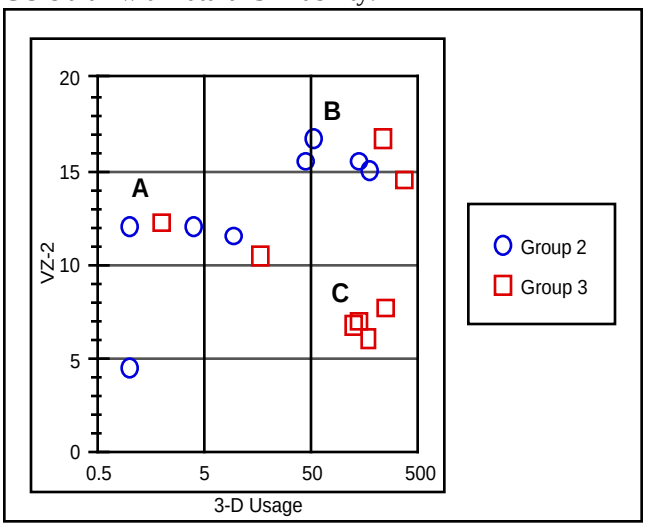

Figure 3. Spatial Aptitude vs. 3-D Usage

Effectiveness measures

The measures used to judge systems in the TREC task were aspect oriented recall and precision. No significant differences were found between any groups or any systems in precision. For aspect oriented recall, ZP outperformed AI, with an average increase of 0.0867 ($p < 0.04$), and AI+ outperformed ZP in recall by 0.0616 ($p < 0.06$). For the samples of participants in our experiment, the differences between our targeted systems and the control system were large, and significant, or nearly significant. If the samples of users are comparable (a tacit assumption of the TREC design) then we should be able to combine our data and

calculate a difference in performance between AI and AI+. Doing this gave a difference in recall of 0.148 ($p < 0.03$ by Tukey's HSD). To verify this result, we ran a direct comparison between the two experimental systems. This found no difference between the two systems at all. Further, there was no overlap between the 95% confidence intervals obtained from the two comparisons. From this we conclude that for the participants selected, there were large differences between our experimental systems and the control system, but the groups of participants were not comparable. In support of this conclusion, the measured traits of the first two experimental groups show large differences in the score for structural visualization ($p < 0.01$). A possible interpretation is that effective usage of a feature-rich GUI depends on strong visualization ability.

We also performed ANOVA to determine if there was any difference in effectiveness between the librarians and general searchers. No differences were found.

Preferred Systems

For the first two groups, we found that librarians preferred ZP over the experimental systems 7 to 1, and general users preferred the experimental systems 6 to 2. For the third group, both librarians and general users preferred AI+ over AI 3 to 1. We find no correlation between system preference and system effectiveness.

CONCLUSIONS

We designed and built two specialized information retrieval systems and tested them. For the groups that the systems were tested on, one was more effective than a control system, and one was less effective. The difference in effectiveness was not found when the two systems were compared directly. Since the sample sizes in the experiment were small it is possible that the groups consisted of individuals who were very consistent within group, but the different groups were not comparable. Structural visualization ability differed strongly between the groups, and may explain some of the difference observed. Usage of a 3-D interface was more strongly influenced by experience with similar interfaces than by structural visualization ability.

We found no correlation between effectiveness in using a system and which system a user preferred. In order to measure system effectiveness, it is necessary to have objective definitions for a given task. We also found no difference in searching effectiveness between experienced searchers (librarians) and novice searchers.

The task performed in our experiment was a straightforward IR task, and allowed the use of standard IR measures for effectiveness. As systems are designed to aid in more open ended, less well defined information exploration tasks, it will be necessary to develop metrics for performance of these tasks.

ACKNOWLEDGMENTS

This material is based on work supported in part by the National Science Foundation under grant number IRI-9619117, and in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and supported in part by Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235.

Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

REFERENCES

[1]Callan, J., Croft, W. B. and Harding, S. "The INQUERY Text Retrieval System" in *DEXA-3: Third International Conference on Database and Expert Systems Applications* (1992)

[2]Croft, W. B. *Organizing and Searching Large Document Collections*, PhD thesis, University of Cambridge, 1979

[3]Ekstrom, R. B., French, J. W., Harman, H. H. and Dermen, D. *Manual for Kit of Factor-Referenced Cognitive Tests 1976* Educational Testing Service, Princeton, New Jersey, 1976

[4]Hearst, Marti A. "Visualization of Term Distribution Information in Full Text Information Access" in *CHI '95 Conference Proceedings*, ACM Press, pp 59-66

[5]Hearst, Marti A. and Pedersen, Jan O. "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results" in *Proceedings of SIGIR '96* (Zurich, Switzerland, August 1996) ACM Press

[6]Swan, Russell and Allan, James "Improving Interactive Information Retrieval Effectiveness with 3-D Graphics", Tech Report TR-96, IR-100, University of Massachusetts Computer Science Department, 1996

[7]van Rijsbergen, C. J. *Information Retrieval* second edition, Butterworths, 1979

[8]Veerasamy, Aravindan and Belkin, Nicholas J. "Evaluation of a Tool for Visualization of Information Retrieval Results", *Proc SIGIR '96*, pp 85-92