Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems

Russell C. Swan and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
http://www.cs.umass.edu/{~swan,~allan}

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 in order to investigate: whether the systems were more effective than a control system, whether experienced users outperformed novices, whether spatial reasoning ability was a good predictor of effective use of 3-D, and whether the systems could be compared indirectly via a control system. Our results show substantial differences in user performance are related to spatial reasoning ability and to a lesser degree other traits. We also obtained markedly different results from the direct and indirect comparisons.

1 Introduction

We are interested in building and evaluating high quality information retrieval and organization tools. We believe that effective use of such tools may require talented users or significant amounts of training. There are many settings where experts in the field are required to spend time learning a tool—e.g., CAD/CAM applications, statistical analysis packages—and the gains from learning the system more than outweigh the time spent learning it. Novice users may find such systems puzzling, but we do not feel that diminishes the value of a targeted system. Further, other researchers are investigating the usefulness of systems for users with little to no searching experience[23, 3].

On the other hand, we have no interest in building systems that are *inherently* difficult to use. Indeed, the better and easier to use a system's underlying design is, the more complexity we can introduce without overburdening the user[21]. For that reason, we are interested in basic issues in interactive computing, among them: how effective are simple system features, how can we compare our various systems, and are there any measures we can use to predict whether a user is likely to be adept at using a particular system or not?

In this study, we investigate exactly those questions. The work was driven by the TREC-6 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. The structure of our experiments was determined to a large extent by the TREC-6 guidelines; they are explained in more detail below.

Because of our interests in targeted systems, we chose to build and evaluate a system that was designed specifically to aid a user with aspect retrieval. The alternative would have been to use a vanilla search engine—perhaps slightly enhanced to look at some specific search technique—for the task; we felt that approach would not sufficiently address our interests. At the same time, we have been investigating 3-D visualizations of document relatedness (clustering), so we chose to create a slightly enhanced version of our system that included a 3-D visualization.

The questions we investigated in the context of this work were:

1. Can we build a system for the aspect retrieval task that is more effective than a basic retrieval system? It is a central hypothesis of our efforts that this is possible.

2. How can we best compare our systems? Can we use a "control" system to compare them indirectly, or must we always compare them directly? There are distinct advantages to comparing via a control (e.g., n rather than n^2 experiments to compare n systems), and an assumption of the TREC-6 track was that doing so would be meaningful.

3. If some users can effectively use our system and others cannot, are there factors that distinguish those users? If so, are they predictive factors that we could have determined in advance? There is suggestive evidence that lead us to hypothesize that verbal fluency would be a good predictor of general performance. We also examined the question of experienced versus novice users by performing half our runs using a group of experienced database searchers: i.e., librarians.

4. Is the 3-D visualization of document relatedness useful? It was our hypothesis that users with strong

To appear in *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998.

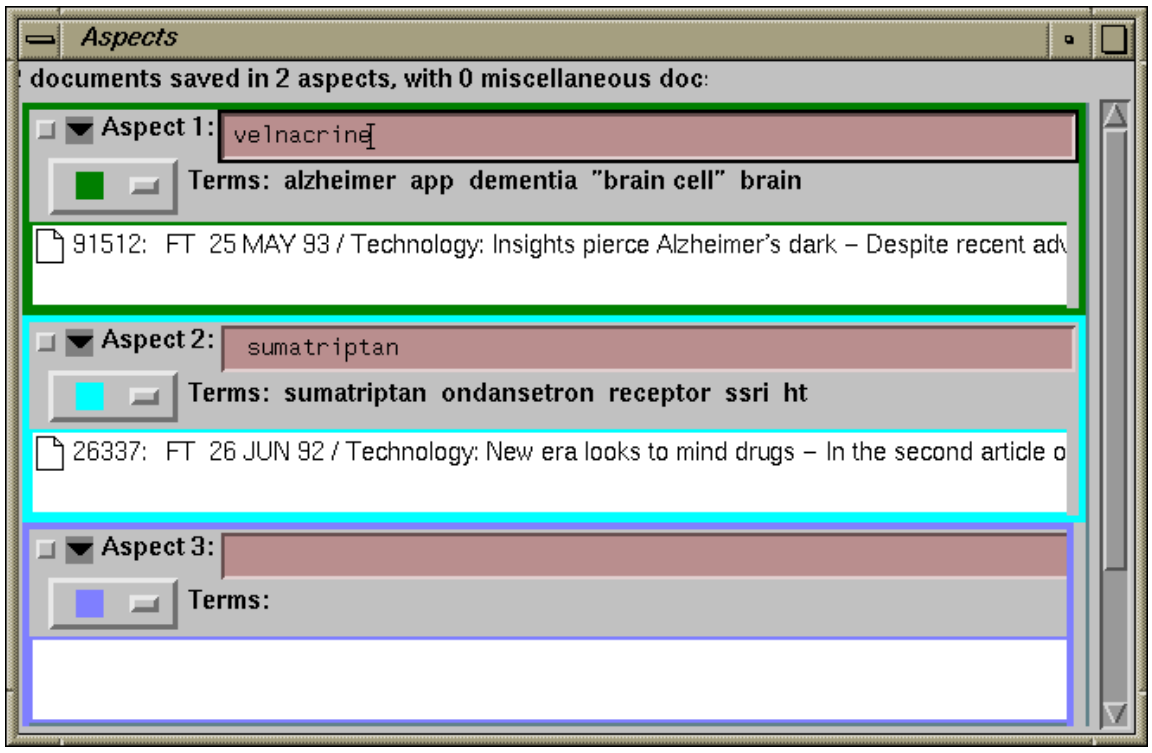

Figure 1: The Aspect Window

spatial reasoning abilities will be able to use the visualization, and that it will increase the number of aspects they can identify.

In Section 2, we describe the three systems that were used as part of this experiment. Section 3 discusses the experimental design and evaluation procedure. In Section 4, we explore how user traits are related to effectiveness (questions 3 and 4 above). Section 5 discusses the issue of system comparison (question 2), and Section 6 covers our ability to generate an effective system for the task (question 1). Section 7 reiterates our conclusions and mentions future work.

2 System

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

1. ZPRISE (ZP) is a basic GUI information retrieval system acquired from NIST. 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, NIST's publicly available search system, modified slightly for the aspect oriented retrieval task (some advanced functionality was removed by 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. For each document in the ranked list, ZPRISE displays the date, the document number, the headline, and a list of terms from the query that were found in that document. When the full text of a document is viewed, query terms contained in the document are highlighted. There is a button for each document on both the ranked list and in the document window; clicking on the button marks the document as being relevant. When a document is first placed in the list the save button is unlabeled. Reading the document causes the label to change to "U". Saving the document causes the label to change to "R".

2.1 Inquery

Our system consisted of the Inquery search engine[5] with a new interface. Our basic user interface has much in common with the ZPRISE interface, including providing visual cues to distinguish between saved/read/unread documents. (Where ZPRISE uses a label on the save button to distinguish state, we write the headline in blue for unread documents and purple for read documents, similar to what Web browsers do, and we place a colored bar before the headline to show which aspect(s) the document contains.) The most significant difference between ZPRISE and our system is that ZPRISE lists the query words contained in a document after the headline, and our system does not.

2.2 Aspect Window

With a basic IR system, an analyst may be able to find the documents containing various aspects, but he or she

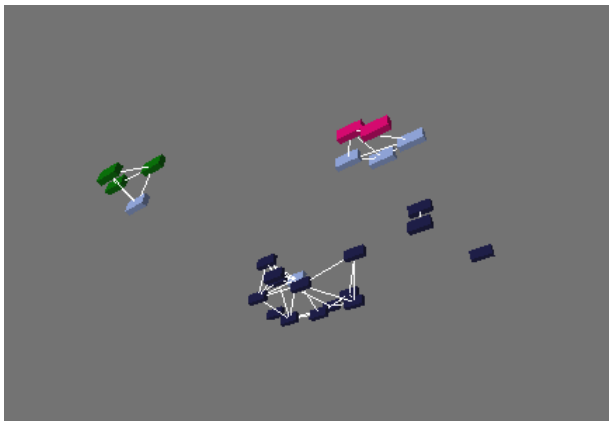

Figure 2: The 3-D Window

has to use another window or a piece of paper to keep track of what has been found already. We implemented an "aspect window" tool to help with this task. The idea is to provide an area where documents on a 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 groups of documents (two aspects) already identified and a third area waiting for the next aspect. The first aspect contains 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".

The purpose of the aspect window is to assist the user in categorizing the information as it is discovered, and to keep an overview of the information discovered so far. In an aspect oriented or briefing type of setting this step is required for the task to be completed properly, but to our knowledge no systems have been built so far which provide any assistance for this task.

2.3 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[12, 23], 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[22] states that relevant documents tend to cluster, and it has been shown to be valid in top-ranked documents[8, 13]. Aspects represent different forms of relevance, and we believe that they will group together within the set of relevant documents.

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 was measured by the sine of the angle between the vectors. That space was collapsed to 3 dimensions for visualization using a spring embedding algorithm (Spring embedding is a force directed placement graph drawing algorithm that generates an approximate solution to a graph layout when the distances between connected nodes are given as constraints. The constraints are modelled as springs[15, 11, 7]). The resulting visualization is similar in style to BEAD[6], differing in a few key aspects: BEAD was used on an entire (though small) corpus, and this display is used only on the retrieved set; the vectors used by BEAD were based on document abstracts rather than the full text.

Documents that are nearby in 3-space are generally nearby in the high dimensional space also (though the spring embedding dimensional reduction occasionally forces unrelated documents to be near one another), 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. An analyst 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. (We have been investigating variations on the visualization that enhance the ability for a user to find new and interesting material [2, 17].)

The three windows—result list, aspect, and 3-D—were 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.

The systems were built so that the use of the Aspect Window is required—there is no other way to mark a document as relevant. The 3-D window is not required and is presented as an alternative to the ranked list. A document can be selected for viewing from either location, and can be opened with a double click in either window. It is possible to navigate through the retrieved documents using only the ranked list, or only the 3-D window, or by going back and forth between the views. We knew that

Trait	Gen mean	StDev	Lib mean	StDev	p
FA-1	27.25	10.93	37.33	11.49	0.05
Education	4.25	2.63	7.42	1.38	0.01
Searching	2.75	1.62	10.67	6.33	0.01
–Library	3.67	1.15	4.83	0.58	0.01
–CD-ROM	2.42	1.38	3.75	1.42	0.05
–Commercial	1.33	0.49	3.33	1.67	0.01

Table 1: Traits showing significant differences between Librarians and General Population. FA-1 is score. Education is years of post secondary education. Searching experience is in years.

usage of the 3-D interface would be highly variable between searchers, so we instrumented the 3-D window to record interactions, recording every time the mouse was clicked on a document in the window, a document was opened, or one of the windows controls (thumb-wheels and sliders) was clicked.

3 Experiment

3.1 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 were recruited for the study and four were placed in each experimental group. All twelve of the librarians had MLS degrees, and several had an additional Masters degree. One had a JD. Ten of the twelve librarians were over forty (the other two were in their twenties). Ten of the librarians were women and two were men.

The general population was recruited by flyers distributed on campus. This group was primarily students (10 of 12 participants). Five were women and seven were men. In most ways this was a very diverse group, ranging from undergraduates to a post doctoral student. However, these people were much younger than the librarians: one participant was in her forties; other than her, the oldest participant was in his thirties. Other traits where there was a significant difference between the librarians and the general population are summarized in Table 1.

3.2 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. All groups participating in the TREC-6 Interactive Track used this experimental design, which is described in greater detail by Lagergren and Over [16].

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. Table 2 shows which systems each group ran.

Group	Block	Population	Control	Exp	Size
1	1	General	ZP	AI	4
	2	Librarian	ZP	AI	4
2	3	General	ZP	AI+	4
	4	Librarian	ZP	AI+	4
3	5	General	AI	AI+	4
	6	Librarian	AI	AI+	4

Table 2: Breakdown of participants

Before the searches, each participant filled out a questionnaire to determine age, education, gender and computer experience, and two psychometric tests[10], a test of verbal fluency (Controlled Associations, test FA-1) and a test for structural visualization (Paper Folding, test VZ-2). We gave each participant a piece of scratch paper before each search, and a short questionnaire after each. Each search had a 20 minute time limit, and the participant was instructed to stop the search if they had not finished in 20 minutes. After all the searches were finished the participant was given a final questionnaire, and then "debriefed". The study was conducted single blind: the participants were not told until the debriefing which system was the control and which was the experimental system.

3.3 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. Six topics were selected by NIST from previous TREC experiments. The documents marked relevant by users were sent to NIST where they were combined with the saved documents from other sites participating in the Interactive Track. The assessors read the documents and developed a list of aspects for each topic, and a mapping between each saved document and the aspect(s) covered, if any. From this, scores of aspectual precision and aspectual recall were obtained for each run. Aspectual precision is the proportion of the saved documents that contained at least one aspect. Aspectual recall is the proportion of identified aspects that are covered by the saved documents. Aspect oriented IR does not entail finding all the documents that mention a topic, as normal IR does, but is instead concerned with finding a set of documents that contains all the relevant information about the topic represented in the corpus.

The first five blocks were run as part of our participation in TREC and were scored by the NIST assessors. We ran block 6 four months later in order to balance our design, and scored the runs using the results from NIST. Eight documents were retrieved by the last block of users that had not been judged by NIST. These were treated as not relevant.

We performed an ANalysis Of VAriance (ANOVA) using MacAnova[19]. More detailed descriptions of our experiment are available elsewhere[4, 1].

4 Traits affecting performance

We are interested in determining if there are any traits influencing searching effectiveness in general, and if there are any traits that lead a user to be more effective with one type of interface than another. In this section we consider the importance of experience and spatial ability,

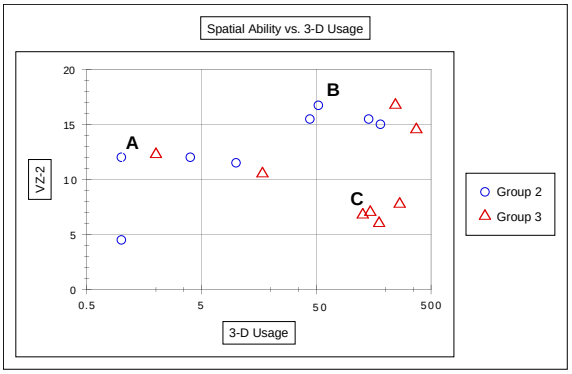

Figure 3: Spatial Ability and 3-D Window Usage for Groups 2 and 3

and look for other traits that are worth measuring. The hypotheses we sought to test were: experienced searchers will be more effective searchers in general than novice searchers; spatial ability will correlate highly with an individual's use of a 3-D interface and their effectiveness with it; and our data will show strong correlations between searching effectiveness and some of the criteria we measured in the psychometric tests and entry questionnaires.

4.1 Novice vs. Experienced Searchers

A distinction is frequently drawn in the IR literature between novice users and experienced users[14]. Librarians are often considered canonical examples of experienced information seekers because they have more information searching experience than the general users, are more educated, and have explicitly studied information and information systems. They differ significantly from our general population in several traits (Table 1), one of which, verbal fluency (FA-1) is believed to correlate strongly with searching effectiveness.

The librarians exhibited different preferences from our general users, with librarians preferring ZP over the experimental system 7 to 1, and our general users preferring the experimental system 6 to 2 ($p < 0.05$). (ZP had an interface very similar to many search engines, and our interface was more novel and very visual). (The librarians in group 3 had the same preferences as the general users, preferring AI+ over AI 3 to 1).

The design of our study allowed us to directly compare the searching effectiveness of the two classes of users. Performing ANOVA on the two classes showed no significant differences in precision, recall, or time taken for both group 1 and group 2. For group 3, the librarians took six minutes less per search on average ($p < 0.03$), but there was no significant difference in average recall or precision.

4.2 Spatial Ability and 3-D Interfaces

Spatial ability is a highly heritable trait that varies greatly among individuals [18]. When a user is confronted with a 3-D interface it is reasonable to expect that their response to it, and effectiveness in using it, correlates with this trait. Before any claims can be made about the usability of 3-D interfaces it is helpful to know where the participants in the study ranked in this trait.

Test VZ-2 measures structural visualization, a form of spatial ability. We used the number of interactions the participant had with the 3-D window during their three searches with AI+ as a measure of usage of the 3-D interface. Figure 3 shows a scatterplot of score on VZ-2 against the number of interactions with 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, and with mouse based interfaces and 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"($p < 0.05$), suggesting that whether or not a person uses a 3-D interface depends more on their familiarity with GUIs than with natural 3-D ability.

4.3 Other Traits

We are interested in determining what other traits are highly predictive of searching effectiveness. One way to accomplish this is by using regression on measured traits to try to build a model of user traits that can replace the user factor blocks. Due to the limited number of users we had we are unable to run a reliable analysis of covariance, but we ran an ANACOVA to look for significant factors that suggest traits that *may* have correlations. On the first group we find significant effects from FA-1, VZ-2, and education, with FA-1 being the most significant. On the second group, we find significant positive correlations for VZ-2 and reported familiarity with mouse based interfaces. Dumais and Schmitt[9] report strong correlations between verbal ability measured by FA-1 and searching effectiveness in a setting without relevance feedback, and a weaker correlation with spatial ability. Our data supply some confirmation that verbal ability and spatial ability are worth further investigation in building user models and are likely to be primary factors.

4.4 Conclusions

Our hypotheses about which traits are useful for predicting behavior and effectiveness are not supported. We see no difference in effectiveness between experienced searchers and novice searchers when we compare librarians against a general academic population. Large differences in effectiveness have been found before[20], but these involved Boolean information retrieval systems and primitive GUIs. Experience and training in a Boolean IR setting may not transfer to a ranked list probabilistic setting, especially with modern GUI systems.

Our hypothesis about who is likely to use a 3-D interface is not supported. The best predictor of who will use an interface element was prior experience with similar elements, not spatial ability. (We discuss usefulness of the 3-D interface, as opposed to use of it, in Section 6.)

Our analysis of the measured traits to look for correlations with searching effectiveness suggested verbal ap-

System	Topic	User	System
AI/ZP Recall	3.45e-12	0.0031	0.0365
AI/ZP Precision	0.0005	NS	NS
AI/ZP Time	0.0021	0.0254	0.0533
AI+/ZP Recall	< 1.0e-15	0.0282	0.0586
AI+/ZP Precision	0.0016	NS	NS
AI+/ZP Time	0.0233	0.0015	NS

Table 3: Significance of Factors

titude and spatial ability, two traits which are already believed to correlate with searching ability.

5 Comparing two Systems

5.1 Comparison via Control

To determine the effectiveness of the two experimental systems we performed ANOVA on group 1 (AI and ZP) and again on group 2 (AI+ and ZP). We treated topic, searcher, and system as factors and precision, recall, and time as dependent variables. We performed the ANOVA with all interactions and found no significant interactions, so we used a main effects model. Table 3 gives the significance figures for each pairing of dependent and independent variables, as determined by the F test in the ANOVA.

Topic is the most significant predictor of recall, precision, and time taken. This is not surprising as it is well known that topic difficulty has a strong influence on IR results. Fortunately the topic effects were quite consistent, and the Latin squares design allowed it to be subtracted out. Without blocking on topics, topic effects would have hidden smaller effects.

User differences were the next most important factor after topic differences. Aspectual recall and elapsed time were both heavily influenced by the searcher. Once again blocking on individual differences is required in order to find system level differences.

System effects were smaller than either topic or user effects, affecting fewer dependent variables and showing far less significance for the variables affected. Three notable system differences were obtained, one significant and two nearly so: ZP outperformed AI in recall by an average increase of 0.0867 ($p < 0.04$), users took an average 104 seconds longer when using AI ($p = 0.06$), and AI+ outperformed ZP in recall by 0.0616 ($p = 0.06$).

The design of the TREC experiment was intended to allow comparisons between different systems by comparing those systems with a common control. We designed our two systems to be identical except for the presence of an additional window in AI+. We felt that if there were a strong difference in effectiveness between the two systems we would know that it was caused by the additional window. If use of a common control allows us to accurately measure system caused differences, we can combine the data for the two groups and perform ANOVA. The ANOVA table for the combined groups 1 and 2 is presented in the Appendix. Significance testing using Tukey's Studentized Range Test shows a difference between AI+ and AI at the 0.03 confidence level, with a ranking of the systems AI+ > ZP > AI, and AI+ outperforming AI in average recall by 0.15, equivalent to finding an additional three aspects out of 20. Since the 3-D window was intended as a recall enhancing device we were encouraged by this result.

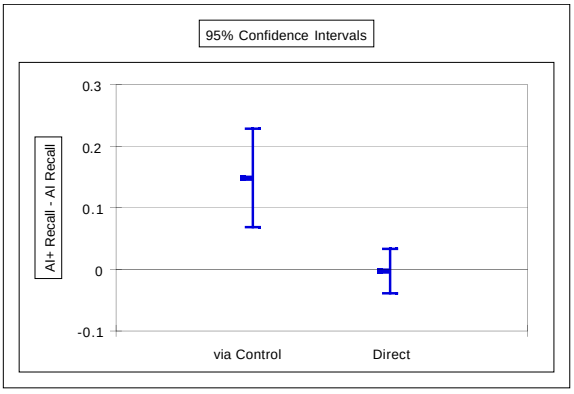

Figure 4: 95% Confidence Intervals for difference between AI+ and AI Recall

5.2 Direct Comparison

In order to confirm this result, and to verify the assumption that different systems could be indirectly compared by comparing them with a common control, we compared the two systems directly in group 3. The ANOVA table is presented in the Appendix. This comparison showed no difference between the two systems in effectiveness. Figure 4 shows the 95% confidence intervals for the difference in mean recall as determined by comparison through the control, and direct comparison. There is no overlap between the two confidence intervals.

System level differences are small compared to differences caused by topics or by users. Ideally, to measure system effects we should hold both topic and user constant across tests. Due to learning effects this is not feasible, and users cannot run the same topic more than once. New users are required for each test, and differences between the sets of users can affect the results. The design of the Interactive Track experiment calls for the use of a common control, the same six topics presented in the same order, and a common Latin Square design to allow indirect comparisons of systems between sites [16]. However, the design only requires four users per system. Small sample sizes can affect experiments in several ways. The most obvious and expected is a reduction in the power of the test—large differences between systems are required in order to obtain statistically significant results. Another problem that can occur with small sample sizes, especially with human subjects, is the possibility of getting highly coherent samples of subjects that are not representative of the population as a whole.

We recruited all our groups the same way, and balanced the distribution of experienced and novice users, but we made no attempt to balance the groups on other traits. We analyzed the characteristics recorded for the different groups of users to see if there were any traits where the groups differed radically. Figure 5 shows the score for Spatial Ability (VZ-2) for the three groups. This distribution of VZ-2 scores for groups one and two has a t-value of 3.707 ($p < 0.01$).

5.3 Interference from Traits

Figure 6 shows the difference in mean recall between the experimental systems and ZP, plotted against VZ-2. (Recall scores were normalized for each topic to have zero mean and unit variance to remove topic effects). Only

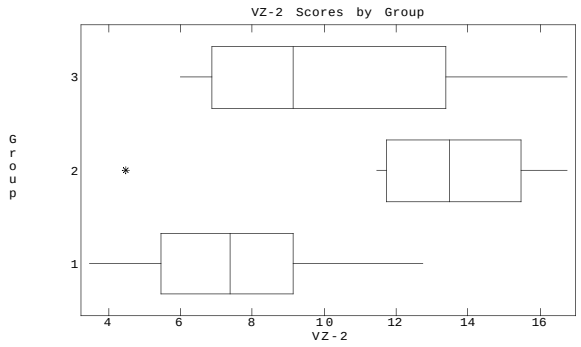

Figure 5: Distribution of Spatial Ability across Groups

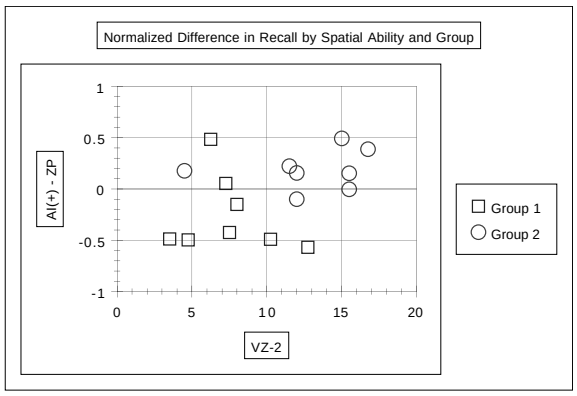

Figure 6: Difference in Experimental/Control Recall and Spatial Ability

two of the eight members of group 1 did better with the experimental system, and only one member of group 2 did worse. Also, only one member of group 1 scored above 11 on VZ-2, and only one member of group 2 scored below that. The difference between systems correlates with score on VZ-2, though not as strongly as it does with group number, with users scoring below 11 doing better with the experimental system 3 out of 8 times, as opposed to users scoring above 11 doing better 5 out of 7 times (with one user doing equally well with both systems).

Not only are there large differences *between* the two groups in VZ-2 score, there are also very small differences *within* each group. A likely explanation for the different results of the two comparisons (via control vs. direct) is that our two systems are essentially identical and *both* of our systems require high visual skills to be effective. The difference in response that we saw is caused by the large difference in average spatial skills between groups, but the differences within groups are too small for the interaction effects to be noticeable.

We had expected to find that users with high spatial skills would find the system with the 3-D window more usable, but we had not expected that result for the basic system. The basic system required the use of Drag and Drop to save documents, and explicitly used a spatial metaphor, where relevant documents had to be dragged to a different window to be saved. This metaphor may be awkward or counterintuitive for users who do not have strong spatial skills. Alternatively, there may be another trait that happens to cluster with spatial ability that explains the difference.

5.4 Conclusion

System effects are small compared to topic and searcher effects. Recall is the only measure that was strongly influenced by system.

The method of comparing different systems by comparing them to a common control is heavily dependent on the users in the study. With the small sample sizes used, not only is power reduced (making it difficult to achieve significant results), but even when significance is obtained it can be an artifact of sampling differences rather than system differences.

6 Targeted Systems

We believe that we can build a system that is effective for a particular task. To that end, we built a system with an aspect window to help with the task, and a 3-D visualization to help the users rapidly find aspects. Were those tools effective?

As seen in Section 5, *for the groups tested*, our first system was less effective in performing this task, while our second system was more effective. The differences in effectiveness are greater than what would be expected by chance, so for these groups, we found strong differences in effectiveness.

The aspect window was intended to help the user with the organization of the information that they had already gathered, but it supplied no information about which document to look at next. The 3-D window provided information about the difference of information content between documents and was intended to be used in place of the ranked list for suggesting which document to view next. We expected that the recall figures for AI would be similar to those for the control, and the recall for AI+ would be higher than the control.

Our AI+ system had two distinct novel interface elements, and we were concerned that users may find a cognitive overload from being presented with both elements. We found a large variation in the amount that participants used the 3-D window, and an extreme response from users about the usefulness of the window, with several participants using the term "worthless" to describe it, and several other participants describing it as natural and intuitive, and wondering why this window is not available on commercial Web search engines.

From the test done with group 3 we see no evidence of effectiveness in the 3-D visualization. We performed a separate ANOVA on each of the three clusters shown in Figure 3 to determine if there were any difference in usefulness of the visualization for groups with different VZ-2 levels or usage levels. The sample sizes were too small to supply statistical significance, but we found that for cluster "B" the systems ranked AI > AI+ > ZP, with AI and AI+ close together. For cluster "C" the systems ranked AI > AI+ (everyone in cluster "C" was in group 3 and did not run ZP). For cluster "A", the set of people who did not use the 3-D window, the rankings were AI+ > AI > ZP. The only set where AI+ outperformed AI was the set that did not use the only feature different in AI+. We conclude that there is no evidence for the effectiveness of this 3-D visualization. It may prove to be useful once people have more experience with it and find it less overwhelming, and as 3-D interfaces become more common, but we have no evidence to support this.

7 Conclusion

In this work, we report on a user study initially undertaken as part of the TREC-6 Interactive Track. In terms of our original goals, we conclude the following:

- We can build a system that is more effective for aspect oriented retrieval than a generic IR system, with one qualification: we were successful with group 2 running AI+ but an equivalent system was less effective at the task with a different group of users. Our conclusion is that for a specific task, *and* a specific group of users, we can build a more effective system. On average (across both groups of users), our systems did slightly worse than the control. We are still confident that we can construct task-specific systems, though we suspect that more experienced users will be needed.
- The TREC goal of comparing two systems indirectly via a common control is not supported by the current experimental design. The current design calls for a minimum of four users per system. We could not obtain consistent results with eight users per system. System specific differences are small compared to topic specific and user specific differences. Since topic effects dominate user effects we must hold topic constant and try to get experimental classes of users that are comparable. In order to get comparable classes of users, we need to know what measurable traits of users are highly predictive of searching effectiveness. When we are capable of building and testing a highly predictive model of user effectiveness we will be able to do cross system comparisons via a control, but our current knowledge of user modeling is inadequate.
- We found a high difference in effectiveness in the use of our systems between two groups of users. These groups differed markedly in their spatial ability, and were otherwise quite homogeneous. We conclude that effectiveness in using direct manipulation UIs is dependent on spatial ability. We also found weak evidence that verbal fluency, spatial ability, and education are factors affecting searcher effectiveness. We found no distinction in searching ability between experienced (librarian) and novice (student) searchers.
- We found no evidence of usefulness for the 3-D visualization. We found that the use of the visualization is better predicted by the users' past experience with GUIs and mouse based interfaces than it is by spatial ability.

As more information becomes available to users and IR systems become more ubiquitous more work will need to be done on the usability and effectiveness of user interface elements for specific tasks. User studies are expensive and time consuming, but without user studies we cannot know what is effective and what is not. Minimizing the number of user studies needed to get valid results will be required in order to find out what works and what does not, but we have seen here that the results of small user studies are not necessarily transitive. In order to be able to conduct indirect system comparisons we need both larger samples and a good model of users. More work needs to be done on first finding the traits that strongly correlate with effectiveness, and then on building accurate predictive models. Without accurate models we cannot design user studies that have reliable results.

Acknowledgments

We would like to thank Don Byrd for his work in setting up and running the user study that is analyzed in this paper. We would also like to thank Victor Lavrenko, Anton Leouski, and Darren Mas for their help in setting up the systems and analyzing the results. We also thank Eva Goldwater for her help with the statistics, and the reviewers for their comments.

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 and also supported in part by United States Patent and Trademark Office and 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 opinions of the authors and do not necessarily reflect those of the sponsor.

References

[1] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu. INQUERY does battle with trec-6. In D. Harman, editor, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. National Institute of Standards and Technology Special Publication, (in press).

[2] J. Allan, A. Leouski, and R. Swan. Interactive Cluster Visualization for Information Retrieval. Technical Report IR-116, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, 1997.

[3] G. Brajnik, S. Mizzaro, and C. Tasso. Evaluating user interfaces to information retrieval systems: A case study on user support. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 128–136, Zurich, 1996. Association for Computing Machinery.

[4] D. Byrd, R. Swan, and J. Allan. TREC-6 interactive track report, part 1: Experimental procedure and initial results. Technical Report IR-117, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, November 1997.

[5] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[6] M. Chalmers and P. Chitson. Bead: Explorations in information visualization. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*, pages 330–337, Copenhagen,Denmark, 1992. ACM.

[7] J. D. Cohen. Drawing graphs to convey proxomoty: An incremental arrangement method. *ACM Transactions on CHI*, pages 197–229, 1997.

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

[9] S. T. Dumais and D. G. Schmitt. Iterative searching in an online database. In *Proceedings of Human Factors Society 35th Annual Meeting*, pages 398–402, 1991.

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

[11] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force directed placement. *Software - Practice and Experience*, 21:1129–1164, 1991.

[12] M. A. Hearst. Visualization of term distribution information in full text information retrieval. In *Human Factors in Computing Systems CHI '95 Conference Proceedings*, pages 59–66, Denver, 1995. Association for Computing Machinery.

[13] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypotheses: Scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, Zurich, 1996. Association for Computing Machinery.

[14] M. Iivonen. Searchers and searchers: Differences between the most and least consistent searchers. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, pages 149–157, Seattle, Washington, July 1995. ACM.

[15] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, 1989.

[16] Eric Lagergren and Paul Over. Comparing interactive information retrieval systems across sites: The trec-6 interactive track matrix experiment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (these proceedings)*, Melbourne, Australia, August 1998. ACM.

[17] A. Leouski and J. Allan. Visual interactions with a multidimensional ranked list. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (these proceedings)*, Melbourne, Australia, August 1998. ACM. Poster presentation.

[18] M. G. McGee. *Human Spatial Abilities.* Praeger Publishers, 1979.

[19] G. W. Oehlert and C. Bingham. Macanova a program for statistical analysis and matrix algebra, 1997. Department of Applied Statistics,University of Minnesota, St. Paul, http://www.stat.umn.edu/~gary/macanova/macanova.home.html.

[20] T. Saracevic and P. Kantor. A Study of Information Seeking and Retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science*, pages 197–216, 1988.

[21] B. Shneiderman, D. Byrd, and W. B. Croft. Clarifying search a user-interface framework for text searches. *D-Lib Magazine*, 1997.

[22] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.

[23] A. Veerasamy and N. J. Belkin. Evaluation of a tool for visualization of information retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 85–92, Zurich, 1996. Association for Computing Machinery.

Appendix

The ANOVA tables comparing recall for AspInquery and AspInquery+ are presented here. Table 4 is the ANOVA for combined groups 1 and 2, and Table 5 is the ANOVA for group 3.

	DF	SS	MS	F	P-value
Constant	1	18.44	18.44	1061.43	0
Topic	5	7.7014	1.5403	88.6350	0
Searcher	15	0.8307	0.0554	3.1877	0.0005
System	2	0.1356	0.0678	3.9041	0.0245
ERROR	73	1.2682	0.0174		

Table 4: ANOVA Table for Indirect Comparison (Combined Groups 1 and 2)

	DF	SS	MS	F	P-value
Constant	1	10.94	10.94	725.20	0
Topic	5	3.8221	0.7644	50.6746	8.5e-15
Searcher	7	0.2817	0.0402	2.66729	0.0259
System	1	0.0004	0.0004	0.0287	0.8666
ERROR	34	0.5129	0.0151		

Table 5: ANOVA Table for Direct Comparison (Group 3)