When Will Information Retrieval Be "Good Enough"?

User Effectiveness As a Function of Retrieval Accuracy

James Allan
allan@cs.umass.edu

Ben Carterette
carteret@cs.umass.edu

Joshua Lewis
jlewis@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01002

ABSTRACT

We describe a user study that examined the relationship between the quality of an Information Retrieval system and the effectiveness of its users in performing a task. The task involves finding answer facets of questions pertaining to a collection of newswire documents over a six month period. We artificially created sets of ranked lists at increasing levels of quality by blending the output of a state-of-the-art retrieval system with truth data created by annotators. Subjects performed the task by using these ranked lists to guide their labeling of answer passages in the retrieved articles. We found that as system accuracy improves, subject time on task and error rate decrease, and the rate of finding new correct answers increases. There is a large intermediary region in which the utility difference is not significant; our results suggest that there is some threshold of accuracy for this task beyond which user utility improves rapidly, but more experiments are needed to examine the area around that threshold closely.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation; H.1.2 [**User/Machine Systems**]: Human factors, Human information processing

General Terms

Performance, Design, Experimentation, Human Factors

Keywords

information retrieval, user study, passage retrieval, performance evaluation

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

1. INTRODUCTION

Information Retrieval research explores a wide range of questions, all somehow connected to the goal of helping a person find useful information in response to a need. The prototypical challenge over almost fifty years of work is improving the accuracy of search systems. The unspoken assumption is that if searchers are given more accurate systems, they will be more successful in finding answers.

That assumption begs a critical question: how much better do search systems need to be for people to appreciate the improvement? A gain of a few percent in mean average precision is not likely to be detectable by the typical user. What if accuracy rose by 50% or even doubled? Even if the user does detect an improvement, does it help with the task at hand? Could it be that search systems are already "good enough" and that more accurate systems will provide at best marginal gains?

In this study, we set out to explore that question. We start with a retrieval task, one that seems broadly interesting to both information seekers and to researchers in the field. We construct an interactive system to help subjects ("users") accomplish the task and deploy state-of-the-art search technology underneath. We measure the time it takes subjects to accomplish the task and how well they actually do completing it, but we do that while varying the quality of the underlying system substantially. When done, we can analyze how search system quality impacts time and effectiveness of the subjects.

Some of these results will be unsurprising. We will show that as system accuracy improves, our subjects find material faster and they make fewer errors of commission (false alarms). However, the actual trend of improvement was unexpected: there is little gain for the subjects until system accuracy has improved a surprisingly large amount. We also found that the subjects' inability to achieve full recall was surprisingly robust: even with highly accurate systems, the subjects on average believed they were done when they had actually achieved only 60% recall.

All data for this study will be made freely available, though the actual documents are proprietary and must be obtained separately. As of publication, the data is available at `http://ciir.cs.umass.edu/downloads`.

The paper proceeds as follows. We start in Section 2 by describing the task that formed the basis of the study. To evaluate system and subject effectiveness, we developed

a collection of topics and relevance judgments, described in Section 3. Our baseline search technology and how we evaluated its quality are discussed in Section 4. Then in Section 5 we describe how we combined truth and system data to create simulated system output at nearly arbitrary accuracy levels. We outline the system used by the subjects, the data we captured, and the measures used to evaluate their effectiveness in Section 6. We present and discuss the results in Section 7, after which we mention prior work and draw conclusions.

2. THE TASK

We believe that two of the core challenges facing information retrieval systems are finding information that is topically relevant and finding novel information within that set [9]. Toward that end, we chose a task that addressed finding multiple facets of an answer to a question. The task embodies passage retrieval, finding relevant information, and finding novel information. It is very similar to and motivated by the TREC interactive track's instance retrieval [15], but requires that the subjects actually identify the facets rather than just find documents containing them. It is similar in spirit to list and definition questions in the TREC question answering task [20]. The idea of variable-length passage retrieval is also explored in the TREC HARD track [2]. The problem of finding novel data is examined in depth in the TREC Novelty track [17].

Specifically, the subject was given a question whose answer had multiple facets (i.e., instances or aspects) and a ranked list of passages that the system chose as most likely to contain those facets. The subject's task was to browse the ranked list of passages and/or their containing documents to find all facets of the answer to the question. Each time they found a facet, the subject highlighted the passage containing the facet and created a "label" that represented the facet. Once they found one facet, they moved on to search for other facets—i.e., novel facets. Subjects could highlight multiple passages for a given facet/label, but doing so gave them neither advantage nor disadvantage in the evaluation.

Ultimately subjects were judged on their ability to find all of the aspects and only the correct aspects and on the amount of time it took them to do so (see Section 6.3.1).

3. CORPUS AND TOPICS

The corpus for this task is a newswire collection from the Linguistic Data Consortium (LDC) comprising stories from the Agence France Press, Associated Press, Los Angeles Times, New York Times, and English-edition Xinhua News. There are 321,590 stories gathered from October 1, 2003 through March 31, 2004, an average of 1,757.3 articles per day. The collection additionally contains documents gathered from April 2004; they were not used for this study.

3.1 Defining Topics

We solicited topic ideas from Information Retrieval researchers. We indexed the corpus using Jakarta Lucene, a Java search engine library that implements a simple vector-space retrieval model [3], and set up a web interface that allowed them to query and browse the corpus. We asked them to write a topic description, including a description of the types of passages that would be considered relevant. A lead annotator, a professional researcher but not an Information Retrieval researcher, vetted topics, tightened up any vagueness, and broadened or narrowed the topic as necessary to ensure that the corpus contained enough relevant material. After obtaining a topic description, we manually created a 1-6 word query and used the system described in Section 4.1 to get a ranked list of 50 unique documents. In some cases the same query was used for a slight variation on a topic.

An example topic is "List the U.S. states or territories that hold caucuses rather than primaries. Labels are names of states or territories." Each state or territory is a facet that might be mentioned multiple times in the retrieved set.

3.2 Annotating Topics

We hired a team of annotators and gave them the ranked list of documents, sorted chronologically, and the complete question. Their job was both to identify relevant passages and to group them by the answer facet they discuss. They used a web interface similar to the one described in section 6.1 to read documents, define labels representing facets (such as "Iowa" for the example topic above), and label passages that support facets. The names of the labels were not important; it was only necessary that the annotator give each facet a different label. Nevertheless, annotators tended to use highly descriptive labels. The annotator was asked to find all passages supporting each facet in the ranked list. A passage could support more than one facet. Passages might overlap.

Two annotators worked separately on each topic. Before beginning annotation, the annotators browsed the ranked list and consulted with each other to agree on what the topic meant and what its facets were. The lead annotator checked both annotations while in progress and offered suggestions for corrections. When both annotators were done with a topic, the lead annotator examined both and chose the one that seemed most correct to use as truth data.

For the example topic above, the selected annotation had 212 passages supporting 19 facets: Alaska, Colorado, Guam, Hawaii, Idaho, Iowa, Kansas, Maine, Michigan, Minnesota, Nevada, New Mexico, North Dakota, Puerto Rico, Utah, Virgin Islands, Washington, Washington D.C., and Wyoming. The other annotator found 164 passages supporting the same 19 facets.

Sixty-two topics were annotated, 45 of which were chosen for use in experiments. The number of facets ranged from 1 to 45, with an average of 13.27 facets per topic, 5.33 passages per aspect, and 66.02 passages per topic.

We calculated interannotator agreement by treating the selected annotation as truth, and calculating the precision and recall of the other annotation at the character level, where precision is the proportion of characters highlighted by the other annotator that were highlighted by the true annotator, and recall is the proportion of characters highlighted by the true annotator that were highlighted by the other annotator. We took the harmonic mean (F1) of these scores as our annotator agreement; by this measure, on average there was 54.2% agreement. This relatively low agreement is one of the reasons we decided to not use precision in evaluating subject accuracy (see Section 6.3.1). Also, the two annotators often identified a different number of answer facets; on average, their number of facets differed by 7.8.

4. BASELINE SYSTEM

Since we were interested in measuring subject effectiveness against ranked lists of varying quality, we used a state-of-the-art retrieval system to get a baseline set of results from which we built those lists.

4.1 Initial system retrieval

Our baseline system was designed to simulate the user task: to retrieve passages relevant to a query and then cluster them for novelty.

Although our truth data used variable-length passages, fixed-length passage retrieval is at least as good as variable-length, and much easier to implement [8, 2, 12]. We therefore divided the corpus into 40-word passages. We ran each query against the corpus and ranked all passages. We borrowed an approach that was successful in the TREC 2004 HARD passage retrieval task: a linear interpolation between a relevance model and a maximum likelihood query model, where the model of term relevance is calculated as:

$$P(w|R) = (1 - \lambda_0)P(w|Q) + \lambda_0(\lambda_1 P(w|R) + (1 - \lambda_1)P(w|C))$$

where $P(w|Q)$ is the query likelihood score of the term with respect to the query, $P(w|C)$ is the background probability of the term given the collection, and $P(w|R)$ is the term probability using relevance models [1]. For this task, both λ_0 and λ_1 were set to 0.5, which gives query likelihood double the weight of the other two models.

For each query, we stepped through the ranked list until 50 unique documents were found. Subjects and annotators only saw passages from those 50 documents. We then clustered the passages using agglomerative threshold clustering [16]; we selected the threshold in each case to get roughly 30 clusters.

4.2 System retrieval quality

Our measure of baseline system accuracy guided what levels of system quality our study should examine. Unfortunately, there is no well-established metric for passage retrieval accuracy. The usual measures for document retrieval do not apply well to passage retrieval, where relevant passages can be of arbitrary length and those lengths are not known to the system. Any given passage might be non-relevant, relevant, or contain both relevant and non-relevant text in arbitrary proportions.

We adopted the binary preference ("bpref") measure that is used for passage retrieval evaluation in the TREC 2004 HARD track [2], because it has been shown both to track mean average precision in ranking system output, and to be a highly stable measure when relevance judgments are sparse [7]. Intuitively, bpref measures the average number of times non-relevant material appears before relevant material. In the case of documents, a bpref of 100% would mean that all relevant documents were ranked above all non-relevant documents. A bpref of 98% would mean that the system ranked 2% of non-relevant material above relevant material.

To apply bpref to passages, we followed the TREC HARD track and measured relevance of characters rather than documents. Every character of a relevant passage is considered, so the measure accurately notes where pieces of the passage were retrieved by a system. If RC is the set of relevant characters and NC is the set of non-relevant characters, and $r(x)$ is the rank of x, then

$$bpref = \frac{1}{|RC|} \sum_{c \in RC} \frac{|\{n \in NC | r(n) > r(c)\}|}{|NC|}$$

That is, it is the average proportion of non-relevant characters that a relevant character outranks.

4.2.1 Passage correspondences

To evaluate clustering accuracy, and also to evaluate annotator agreement and improve ranked lists towards truth, we drew up a set of correspondences between system passages and true passages based on their overlap. We used a broad definition of overlap, such that the algorithm for associating system passages with true passages was as follows:

1. For each true passage, find all system passages that overlap with it at all.
2. If a system passage overlaps with more than one true passage, associate it with only the true passage with the largest number of overlapping characters. Break ties randomly.
3. If a system passage overlaps with no true passages, record it as corresponding to a null passage.

4.2.2 Evaluating clustering accuracy

To evaluate clustering accuracy, we first created a similar set of correspondences between the true clusters and the system clusters. We did this by scoring each system cluster against each true cluster based on how many passages it had in common with that true cluster, and assigning the highest scoring such cluster to the true cluster. Clusters were associated one-to-one, so some remained unassigned. We then relabeled the system clusters with their corresponding true cluster names. A system passage's correct cluster was the cluster of its corresponding true passage, or a special cluster used to collect all non-relevant system passages. So the clustering accuracy of a ranked list is just the proportion of passages in the list that are correctly labeled.

4.2.3 Variance in baseline quality

Baseline passage retrieval quality varied highly across topics, from 0.333 to 0.818. We discarded several topics from use in the study because their baseline scores were too high to be significantly different from the true data. There was no direct relationship between the nature of the query and its baseline accuracy. This demonstrates that we don't have an intuitive grasp on what makes a given query "hard" or "easy" for a system, a problem that makes it hard to identify how best to improve systems [6]. Such intuition would greatly help guide research in retrieval, since for "easy" queries, our state-of-the-art may have very high user utility already, and we would rather concentrate on "hard" queries.

We identified as "hard" queries those which had a baseline bpref of less than 0.60, and identified all better-scoring queries as "easy." The average bpref across all topics was 0.62; the average bpref of "hard" topics was 0.47 and the average of "easy" topics 0.72. Note that our definition of "easy" and "hard" is system-dependent; other systems may have a very different range of performance across these topics.

System	bpref scores				
	min	max	avg	"easy" topics	"hard" topics
Us	0.33	0.82	0.62	0.72	0.47
Lucene	0.06	0.95	0.48	0.50	0.38
Desktop	0.04	0.47	0.22	0.24	0.18

Table 1: Baseline systems compared.

4.2.4 *Our system versus other systems*

To determine whether our retrieval system was a reasonable choice as a baseline, we compared its results to two search systems: a widely available desktop search tool, and Lucene, the Java search engine library mentioned earlier. Note that neither system has been tuned for this task, and results might be different if they were, but both are commonly used general search tools that might represent an alternate baseline than our research system. We evaluated the tools on each of the topics that we used for the study, using the original query strings.

The desktop search tool performed at a much lower rate than our system; this may be because its results list was very short, and the tool had to be coerced into providing sufficient results to be evaluated. Our system outscored the tool on all queries. The average bpref over all topics was 0.220; the range was from 0.040 to 0.469. The average bpref over the topics our system found "easy" was 0.241; the average over our "hard" topics was 0.181.

The Lucene search engine did a much better job than the desktop tool, but its results had higher variance than ours; its bpref scores ranged from 0.058 to 0.953. Although it outscored our system on several topics, on average it did not perform as well: its average bpref over all topics was 0.479. Its average over our "hard" topics was 0.377; over our "easy" topics, 0.504.

Our system substantially outperformed both tools on average; we conclude that it is reasonable to call our results a state-of-the-art baseline for this task.

5. TEST DATA SETS FOR SUBJECTS

The goal of this study was to examine increasing levels of system quality from the current state of the art to the theoretical perfect system. Given the results of our system retrieval, and the truth data created by annotators, we artificially created ranked lists at intermediate levels of accuracy by improving the baseline system output towards truth. This kind of study, in which human-generated approximations of computer output are used so that the quality of the data is entirely known and controlled, is sometimes called a "Wizard of Oz" study. This technique is commonly used in the speech recognition and natural language interface communities to improve system robustness [10, 5, 13], and we have adopted it to examine robustness in Information Retrieval.

It is not possible, of course, to know which types of problems IR technology will correct as it moves forward, and our goal was not to propose a plan for improvement. Instead, we *randomly* improved system output. We created a list of changes necessary to transform baseline system output into perfect output. We then randomly selected one change to make, measured the accuracy of the new ranked list, and stopped if it met or exceeded the target accuracy. We seeded the random number generator with the same number each time, so higher-accuracy blended ranked lists recapitulated the lower-accuracy blended ranked lists during the improvement process.

We had to improve both highlighting accuracy and clustering accuracy.

5.1 Improving passage highlighting accuracy

To find the ways in which the system's output differed from truth, we first found the correspondences between the system passages and the true passages, as described in Section 4.2.1.

There were several ways in which the system's output could differ from truth:

1. A true passage might not be highlighted by the system at all;
2. A system passage might not correspond in any way to a true passage;
3. A system passage might overlap a truth passage, extending to one side, the other, or both;
4. A system passage might overlap a truth passage, but omit some of the truth passage on one side, the other, or both;
5. A true passage might be highlighted partially by multiple system passages.

We fixed one such problem at each step, and then evaluated the new ranked list using the binary preference measure described above. We stopped when we met or exceeded the target accuracy. We blended hard topics (those with a low baseline bpref) to bprefs 50, 60, 70, 80, 90 and easy topics (those with high baseline bpref) to bprefs 80, 85, 90, 93, 98.

5.2 Improving cluster accuracy

As before, we moved the system output toward truth by randomly selecting an incorrectly clustered passage and correcting it, and repeating until we had reached or exceeded a desired level of accuracy. Unfortunately, we discovered partway through the study that clustering accuracy seemed to have no impact on subject effectiveness; subjects apparently ignored the clustering information provided. Other research has supported the finding that clustering may not be useful for QA-like tasks [22]. For that reason, we do not discuss the impact of clustering further.

6. DESCRIPTION OF THE EXPERIMENT

We located subjects by distributing flyers and posting advertisements. We had 33 respondents, who went through an interview and training session designed to both screen unenthusiastic subjects and allow subjects to answer a simple example topic and decide if they wanted to continue. After hiring but before starting any task, there was a second training session to re-familiarize subjects with the system and compare their performance on a second example topic to the truth.

We arranged the 45 topics into 9 problem sets. A problem set consisted of five topics, each represented at five different retrieval accuracy levels. Subjects were asked to complete the five topics in a problem set before moving on to a new problem set. Every five subjects to start a problem set completed their five topics in a different order (determined by a

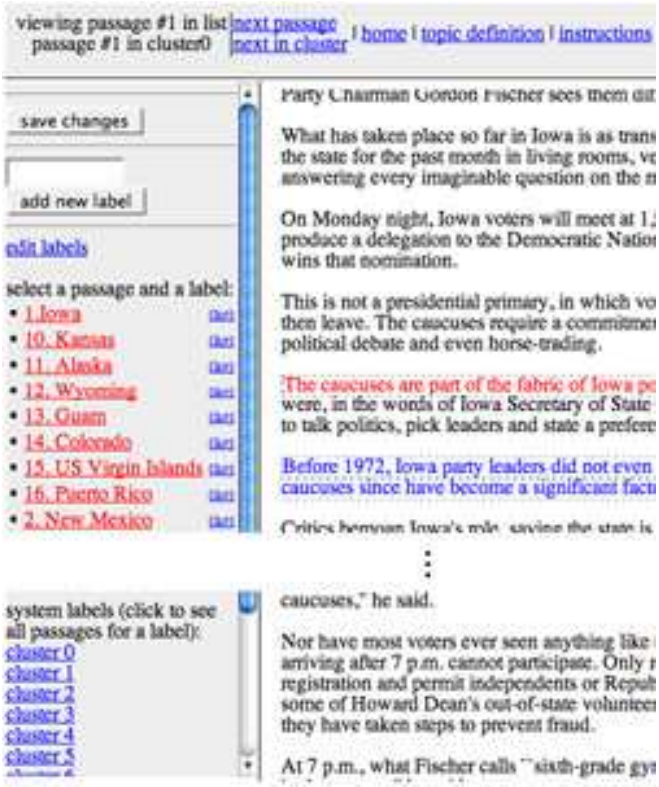

Figure 1: Annotation interface used by subjects to complete tasks. The document text appears in a window on the right. Subject-defined labels are listed on the left in red. System clusters are listed in blue underneath subject labels. Links at the top are used for navigation and information.

Latin square) and at different retrieval accuracy levels. That is, if five subjects completed problem set 1, we would have data for all five topics in the problem set at all five accuracy levels. No subject did any topic more than once.

The subject's goal was to find and highlight at least one passage representing every facet of the topic. They were given unbounded time, but told that it would take about 8 hours, or two sessions, per problem set.

Our hypothesis was that if the system was very bad, the subject would have to spend more time reading the documents to look for relevant passages, while if the system was very good, the subject could simply click on a relevant passage, highlight it in the document, and move on to the next passage, doing very little reading. For the intermediate accuracy levels, the subject perhaps would skim documents at speeds varying according to the accuracy level.

6.1 User interface

The user interface was web-based, with dynamic HTML on the client side and a Perl backend on the server side.

Upon beginning a session, the subject was given the list of questions for the topics in the problem set he/she was working on (or for a new problem set if necessary). The questions were the same as those given to the annotators, with any specific answers that the annotators had been told about removed. The subject was given a random username for each topic; usernames map to (subject, topic) pairs. On logging in, the subject was presented with a ranked list of passages. They were told that the ranked list was the output of a retrieval system. The subject had no knowledge of the true accuracy of the list; they were asked to highlight novel passages until they felt they had exhausted the topic.

Clicking a document title brought the subject to the annotation interface shown in Figure 1, in which they could read the document. In the annotation interface, system-retrieved passages were colored blue. The highlighted text in Figure 1 beginning "Before 1972..." is an example. The subject created new labels by typing a label name into the text box shown in the figure. Labels were listed to the left. To highlight a relevant passage, the subject selected the text with the mouse and clicked the label name. This colored the passage red, shown in Figure 1 with the text beginning "The caucuses...". Any passage could be selected and labeled multiple times. Passages were allowed to overlap.

There was also the cluster feature. System clusters were listed underneath subject labels with non-descriptive names "cluster 0," "cluster 1," etc. Subjects could click on a cluster name to open a window from which they could view and navigate between all passages in that cluster. Subjects could also click on the "(list)" link next to their label to see a list of passages they assigned to that label. It seems that few subjects made use of this facility, however.

6.2 Data captured

Each passage the subject highlighted was saved to disk, with the name of the document, the passage text, the label, the start offset (relative to the HTML-coded document), and the length. In addition, we recorded login and logout times, coffee breaks, transitions between documents, saves to disk, and miscellaneous other information.

In addition to the interface logs, the server logs recorded each time any page relating to the interface was served. Subjects could be matched by IP address.

6.3 Measuring subject effectiveness

Recall that a subject's final answer to a question consists of a set of clusters, representing facets of the answer, each of which contain at least one highlighted passage representing the content of the cluster. Subjects did not know how many facets the correct answer contained, though for some questions they were given an approximate target, e.g. "There are likely to be over a dozen."

Many user studies done in the domain of Information Retrieval examine information visualization techniques to improve user effectiveness, and the majority of these use time on task as their primary metric (for example, [11, 19, 21]). Accuracy measures, such as precision and recall, have been used in some studies, including the TREC Interactive Track [14, 18, 15]. We were interested in accuracy as well as speed, and used the metrics discussed below.

6.3.1 Precision, recall, and error

We can use the traditional measures of precision and recall to evaluate how well the subject "retrieved" facets. In this case, precision is the proportion of the subject's facets that are correct, and recall is the proportion of the correct facets that the subject found. These numbers were calculated by making correspondences between the subject's highlighted

Passage accuracy	Number trials (subject-topic)	Unique topics	Unique subjects
50%	32	8	19
60%	46	13	23
70%	61	18	23
80%	143	40	24
85%	111	32	25
90%	132	40	25
93%	91	28	23
98%	85	23	22

Table 2: Data collected from subject sessions.

passages and the correct passages, as per the baseline system evaluation. A subject found a facet if they highlighted a passage that was part of that facet in the truth data; once a subject's facet label was associated with a true facet label, no other true label could be associated with that subject label or vice versa.

Given the task, we ended up using recall alone as our metric. Recall captures how well the subject found the facets found by the annotator; extra facets marked by the subject may represent highlighting errors but usually reflect annotator disagreement.

We expected recall to mostly be close to perfect, since the answers were always findable no matter what the system quality, but that was not the case. Supporting the findings of Blair and Maron [4], subjects were generally poor at deciding when the task is complete.

In evaluating error, miss errors are captured by failure of recall to reach 100%. We also explicitly measured the number of false alarms—i.e., nonrevelant passages that the subject highlighted. Usually, false alarm rate would be defined as the number of marked non-relevant passages over the total number of non-relevant passages. Because passages are of variable length, it is unclear how to define the total number of non-relevant passages, so instead we report the actual count of false alarm passages. This number should be taken with a grain of salt, since it might simply reflect disagreement.

6.3.2 *Time and Recall per Time*

We measured the subject's overall time spent on each topic, measured in minutes; if a subject worked on a question in multiple sessions, the times were summed without counting the break between. In addition, because recall scores were rarely close to 100%, we used a subject's recall divided by their time to normalize how much information was found in the time spent.

7. RESULTS AND ANALYSIS

Of the 33 subjects hired, 18 completed five or more problem sets. Eleven of those completed nine problem sets (all 45 topics). Six subjects were removed from the analysis due to an incomplete understanding of the task.

Table 2 shows the number of (subject,topic) pairs we obtained data from at each passage accuracy level. The data sparseness at the low levels is due to most of our topics having a high retrieval baseline, which we did not expect. We had approximately the same number of subjects at each accuracy level, so we expect that differences between subjects

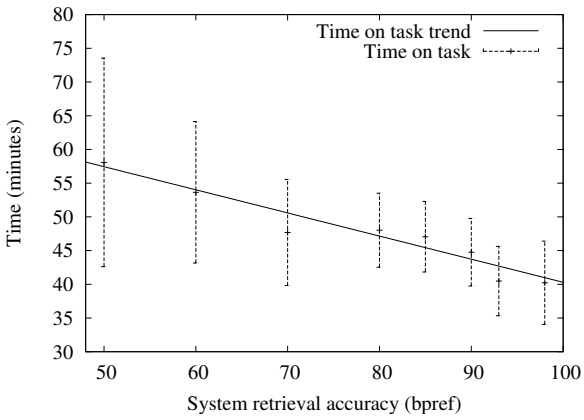

Figure 2: Average time vs. system accuracy.

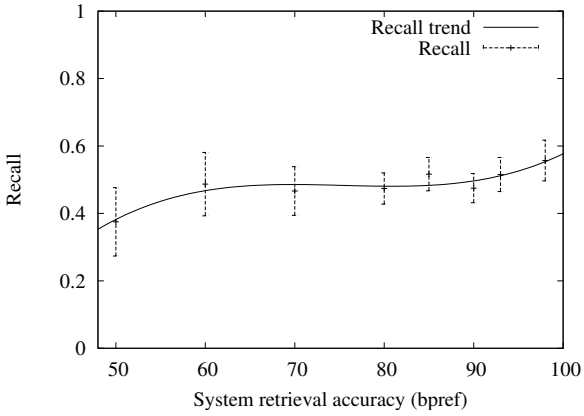

Figure 3: Recall vs. system accuracy

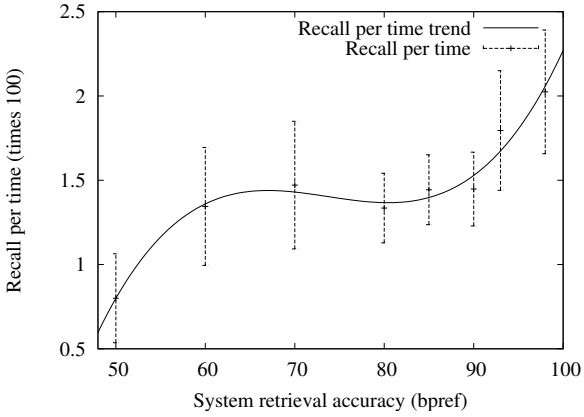

Figure 4: Recall per time vs. system accuracy

will average out. The number of topics at each accuracy level varies quite a bit, which complicates analysis. It is not entirely clear whether the difference in results at, say, 50 and 98 is due to the difference in accuracy or to the difference in topics at both levels. However, when we break the data into

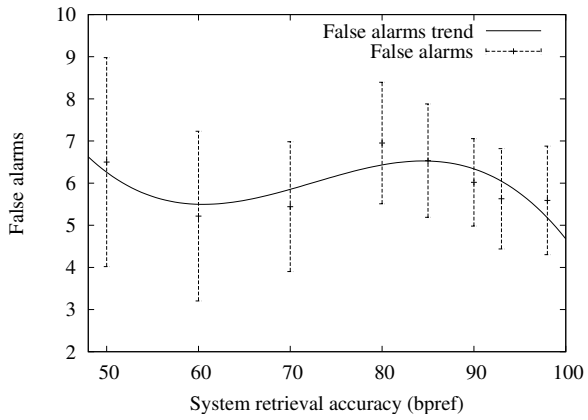

Figure 5: False alarms vs. system accuracy

two groups ("easy" topics with baseline bpref greater than 60 and "hard" topics with baseline bpref less than 60), we observe the same trends in both groups that we see in the combined data. This lends credence to our claims that the trends are real.

We averaged each evaluation measure over subjects and topics at each accuracy level. To test significance, we performed Student t-tests between all pairs of accuracy levels (e.g. 50 and 60, 50 and 70, ...). In general, the significance numbers should be taken with a grain of salt—as stated earlier, sources of variance are not clear.

Statistical analysis might be influenced by factors such as subjects getting better due to "learning the system", subjects having prior knowledge of a topic, or subjects getting tired and less effective towards the end of a session. We hope that randomization of topic presentation mitigates these factors.

Figure 2 shows a decrease in subject time on task as system retrieval accuracy improves. At passage accuracy 50, average time on task was about 58 minutes, with high variance. At passage accuracy 98, average time on task was about 40 minutes, a 31% decrease. The difference in time is significant at $\alpha = 0.95$ between 50 and 93, 60 and 90, 80 and 93, 85 and 93, and 93 and 98. At $\alpha = 0.9$ we have significance between 50 and 80, which suggests that with more data we could claim greater significance.

Figure 3 shows an increase in recall from 0.38 to 0.56 as system accuracy improves. Several topics gave subjects a "hint" about how many facets they should be able to find, but even on these topics average recall was fairly low. Difference in recall is significant at $\alpha = 0.95$ between 50 and 80 and above, and between 90 and 98. At $\alpha = 0.9$, the difference is significant between 50 and 60.

Figure 4 shows recall per time (multiplied by 100). As the figure shows, it is fairly constant from passage accuracy levels 60 to 90, but increases steeply from 50 to 60 and from 90 to 98. At $\alpha = 0.95$, the difference is significant between 50 and 60 (and above), between 90 and 93, and between 98 and 90 (and below).

Figure 5 shows a slight decrease in "errors" (disagreements). In general, it decreases where Figure 4 increases. None of the differences are significant at $\alpha = 0.95$.

When judging subjects against the second, unused annotation for each topic, the trends were generally the same. The most notable difference was slightly lower recall at each accuracy level.

It is not surprising that time on task decreases as system accuracy increases, but it is interesting to note that the variance in time is substantially smaller with better system output. That is, expected performance is likely to be more predictable with better systems—to within 10 minutes in our study—whereas the variation is over 30 minutes for current technology.

The failure of subjects to improve recall (even when guided toward the right target) reinforces the need for work on interfaces or retrieval techniques that help people understand a broad topic better. For queries where there is only one answer (e.g., "what is the home page of XYZ corporation?") this issue is unimportant. But for queries where better (if not perfect) recall is important—e.g., medical advice, intelligence analysis, or comparison shopping—there is great value in enabling people to find more and different information. It is surprising that even with incredibly accurate retrieval (in comparison with the state of the art) our average subject did not find more than 60% of the facets of relevance.

Possibly the most interesting graph of results is Figure 4, showing how time per recalled facet changes with system accuracy. Recall that we had two classes of topics. The easy topics resulted in system output in the middle of the curve's "ledge" from 60-90% accuracy. If the trends of the study are accurate, this suggests that it will be very difficult to improve a user's experience for those topics: only by improving accuracy 25-40% relative will there be gains. For the more difficult queries, the gains in user effectiveness will appear much more quickly, but will then stop at the ledge. This suggests that research aimed at quick overall improvements should focus on challenging queries, and that substantial gains will otherwise only be reached with significant investment or new approaches to these tasks. Since for this task recall is strongly related to novelty (it is facet recall), alternative ways to tackle finding novelty may be very important.

8. CONCLUSIONS

Our experiments verify our expectation: Improving retrieval systems results in greater user effectiveness. With better retrieval, users can work faster, find more facets, and make fewer errors. Our experiments suggest that improving "hard" topics is about as useful as improving "easy" topics; that is, the average improvement in user effectiveness on easy topics is approximately the same as the average improvement in user effectiveness on hard topics when retrieval performance is increased over the ranges we used.

The shape of the graphs suggest that an improvement from a bpref of 50 to 60 provides the biggest benefit in user effectiveness for hard topics, while it will take a larger improvement to provide an equivalent benefit for easy topics. This suggests that the IR community should focus on improving performance on difficult topics. We define "hard" to be that which our system performs poorly on; we need a more rigorous and less circular definition. A suggestion for future work, then, is some way to differentiate hard topics from easy, and to more precisely evaluate performance differences on hard topics.

All of this suggests that there is still quite a bit of work to be done to improve retrieval systems before users see a clear

difference. The fact that users *will* experience a difference is important: although we intuitively feel that search engines generally give good results, we can see that there is room for significant improvement.

It is important to note that to the extent they are meaningful, these results may only apply to the task we selected for this study. Simple ranked retrieval with no concern for novelty may have completely different utility characteristics. Requiring that subjects find *all* passages relevant to a facet (i.e., the annotators' task) would presumably cause different results. Tasks that require subjects to work much further down the ranked list would have lower overall scores and possibly different implications. Nonetheless, although we are cautious in how we interpret the information, we would like to think that this study is more broadly applicable. It clearly points out problems with facet recall, and by extension recall in general. It suggests that advances in retrieval effectiveness may be difficult for users to appreciate until they are pronounced improvements. It also indicates that better systems may result in more predictable performance.

9. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

10. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, M. Smucker, T. Strohman, H. Turtle, and C. Wade. UMass at TREC 2004: Notebook. In E. Voorhees, editor, *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, pages 657–670, 2004. Available online at `http://trec.nist.gov`.

[2] J. Allan. HARD track overview in TREC 2004 (notebook), high accuracy retrieval from documents. In E. Voorhees, editor, *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, pages 226–235, 2004. Available online at `http://trec.nist.gov`.

[3] Apache Software Foundation. Jakarta Lucene. Can be found at `http://jakarta.apache.org/lucene/`.

[4] D. C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299, 1985.

[5] S. Boyce and A. Gorin. User interface issues for natural spoken dialog systems. In *Proceedings of the International Symposium on Spoken Dialogue 96*, pages 65–68, 1996.

[6] C. Buckley. Why current IR engines fail. *Proceedings of the 27th annual ACM SIGIR conference*, pages 584–585, 2004.

[7] C. Buckley and E. Voorhees. Retrieval evaluation with incomplete information. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[8] J. P. Callan. Passage-level evidence in document retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310, 1994.

[9] Z. Chen and Y. Xu. User-oriented relevance judgment: A conceptual model. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, page 101, 2005. Abstract is in proceedings; full paper in accompanying CD-ROM.

[10] N. Dahlback, A. Jonsson, and L. Ahrenberg. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pages 193–200, 1993.

[11] O. Drori and N. Alon. Using documents classification for displaying search results list. Technical Report 2002-46, Leibniz Center for Research in Computer Science, Hebrew University, Jerusalem, Israel, 2002.

[12] M. Kaszkiel and J. Zobel. Passage retrieval revisited. *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 1997.

[13] A. Kehler, J. Martin, A. Cheyer, L. Julia, J. Hobbs, and J. Bear. On representing salience and reference in multimodal human-computer interaction. In *AAAI '98 (Representations for Multi-Model Human Computer Interaction)*, pages 33–39, 1998.

[14] J. Koenemann and N. J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *CHI*, pages 205–212, 1996.

[15] P. Over. TREC-6 interactive track report. In E. Voorhees and D. Harman, editors, *The Sixth Text Retrieval Conference (TREC 6)*, pages 73–82, 1998. NIST Special Publication 500-240.

[16] G. Salton. *Automatic text processing: the transformation and analysis and and retrieval of information by computer*. Addison-Wesley, 1989.

[17] I. Soboroff. Overview of the TREC 2004 novelty track. In E. Voorhees, editor, *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, pages 57–70, 2004.

[18] R. Swan and J. Allan. Improving interactive information retrieval effectiveness with 3-d graphics. Technical Report IR-100, University of Massachusetts, Amherst, Center for Intelligent Information Retrieval, 1996.

[19] E. Tanin, A. Lotem, I. Haddadin, B. Shneiderman, C. Plaisant, and L. Slaughter. Facilitating network data exploration with query previews: A study of user performance and preference. Technical Report CS-TR-3879, University of Maryland, College Park, 1998.

[20] E. Voorhees. Overview of the TREC 2003 question answering track. In E. Voorhees, editor, *The Twelfth Text Retrieval Conference (TREC 2003)*, pages 73–82, 2003. NIST Special Publication 500-255.

[21] W. J. Weiland and B. Shneiderman. A graphical query interface based on aggregation/generalization hierarchies. *Information Systems*, 18(4):215–232, 1993.

[22] M. Wu, M. Fuller, and R. Wilkinson. Searcher performance in question answering. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 375–381, 2001.