

Impact of Retrieval Precision on Perceived Difficulty and Other User Measures

Mark D. Smucker
Department of Management Sciences
University of Waterloo
msmucker@uwaterloo.ca

Chandra Prakash Jethani
David R. Cheriton School of Computer Science
University of Waterloo
cpjethan@cs.uwaterloo.ca

ABSTRACT

When creating interactive retrieval systems, we want to reduce the perceived difficulty of finding relevant documents. We conducted a user study that controlled retrieval precision. We found that a higher retrieval precision caused a reduction in perceived difficulty compared to a lower retrieval precision. We also found that higher precision increases enjoyment and has some influence on ability to concentrate, but we found no evidence that precision keeps the user engaged vs. bored with the search.

1. INTRODUCTION

In this paper, we examine whether or not the user perceives any differences in the search experience given a change in precision. In other words, does retrieval precision affect how a user feels about a search for relevant documents? We certainly expect the topic to have significant impact on the degree to which a search seems difficult or enjoyable, but does precision?

To answer this question, we utilize data collected as part of larger user study that we conducted to examine the relationship between retrieval precision and human performance [5]. In that study, we examined the effect of two levels of precision on the performance of users. We looked at search results with a uniform precision at rank N of 0.3 and 0.6. Users were to find as many relevant documents as possible within 10 minutes. We asked the users to work quickly and to balance their speed with accuracy. We found that precision is strongly related to performance for the interfaces and tasks of our study.

While the two levels of precision resulted in different levels of human performance, did the retrieval precision affect the users' perception of search difficulty? We can improve user performance as measured by some metric of our choosing, but if users do not notice this measured performance improvement, then we may need to reexamine our conception of performance. We found that:

- Retrieval precision has a statistically significant effect on the perceived difficulty of finding relevant documents ($p = 0.006$) as well on the enjoyability of the search experience ($p = 0.016$).
- The user's ability to concentrate is somewhat impacted by retrieval precision ($p = 0.079$).

- The mood of the user (bored or engaged) is not affected by retrieval precision ($p = 0.341$).

These results add support to results of Bailey, Kelly, and Gyllstrom [3] who found that their estimate of topic difficulty correlated with perceived difficulty of finding relevant documents. Bailey et al. estimated a topic's difficulty for users by using a collection of existing user queries for the topic and measuring the average nDCG of these queries.

There are many ways to describe topic difficulty. For example, a topic could be hard for users to understand and distinguish relevant from non-relevant documents. Conversely, a topic could be easy to understand, but the topic could require careful inspection of documents for relevance if there are many requirements attached to what makes a document relevant. Users can vary greatly in their familiarity of a topic, and these differences in familiarity could affect the user's perception of the topic difficulty. Another notion of topic difficulty may be to consider easier topics to be those topics that allow more relevant documents to be found in a given amount of time than harder topics.

Rather than attempt to define topic difficulty and determine its affect on various user measures, we control the precision of retrieval results for a set of topics. Many of the ways to describe topic difficulty are likely independent of the precision of the results the user is examining. Our primary contribution in this paper is that we show evidence that precision causes changes in perceived difficulty. While we believe that Bailey et al. also found that that users perceive it easier to find relevant documents the higher the retrieval precision, they did not directly control the retrieval precision nor did they vary precision across a set of topics.

2. METHODS AND MATERIALS

In this section, we briefly describe our user study. Details can be found in our earlier publication [5].

We conducted a two phase user study. 48 users participated in the study. The same users participated in both phases of the study. For each phase, users completed 4 search tasks. A task corresponded to searching for documents relevant to a given TREC topic. Each task took 10 minutes. In total, we used 8 topics from the 2005 TREC Robust track, which are shown in Table 1. The 2005 Robust track used the AQUAINT collection of 1,033,461 newswire documents.

Each phase used a different user interface. Figure 1 shows the user interfaces. In the first phase, users judged the relevance of document summaries and full documents. Users saw one summary or document at a time and had to judge

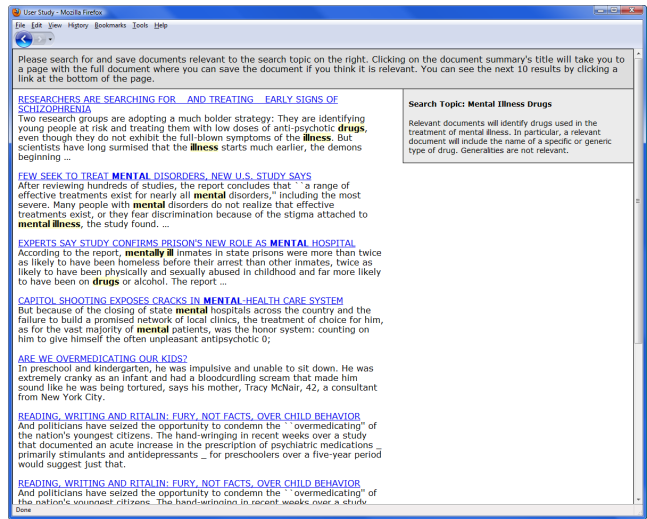
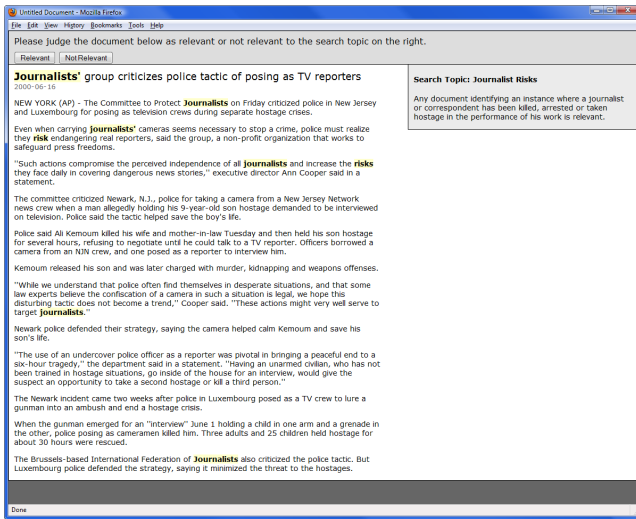


Figure 1: The left screenshot shows the user interface (UI) for phase 1 with a full document. Participants in phase 1 also judged document summaries in the same way. The right screenshot shows the phase 2 UI with query-biased document summaries shown. Clicking on a summary took the user to a page with the full document. This page, similar to the phase 1 document judging UI on the left, allowed the user to save the document as relevant, but did not require a relevance judgment be made.

| Number | Topic Title | Relevant |
|--------|------------------------------|----------|
| 310 | Radio Waves and Brain Cancer | 65 |
| 336 | Black Bear Attacks | 42 |
| 362 | Human Smuggling | 175 |
| 367 | Piracy | 95 |
| 383 | Mental Illness Drugs | 137 |
| 426 | Law Enforcement, Dogs | 177 |
| 427 | UV Damage, Eyes | 58 |
| 436 | Railway Accidents | 356 |

Table 1: Topics used in the study and the number of NIST relevant documents for each topic.

the document to see the next document. Summaries and documents alternated.

In the second phase, the user interface was similar to today's web search engines that display 10 query-biased summaries in response to a user's search query. Clicking on a summary showed the user the full document, and the user could choose to save the document if it was relevant. If the user believed the document was non-relevant, or did not want to take further action on this document, the user would click the web browser's back button to return to the search result summaries. Query reformulation was not possible. While phase 1 restricted users to making judgments in the order of the search results, phase 2 made no such restriction.

As part of the study, the users answered a questionnaire after each task. For the questionnaire, we used the same 4 questions as used by Bailey et al. [3] in their work:

1. How difficult was it to find relevant documents about this topic?
2. How would you rate your experience searching for information about this topic?

Post-Task Questionnaire

Search Topic : Black Bear Attacks

A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior. It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

1. How difficult was it to find relevant documents about this topic?
 - Very Difficult
 - Difficult
 - Neutral
 - Easy
 - Very Easy
2. How would you rate your experience searching for information about this topic?
 - Very Unenjoyable
 - Unenjoyable
 - Neutral
 - Enjoyable
 - Very Enjoyable
3. How would you rate your mood while you searched?
 - Very Bored
 - Bored
 - Neutral
 - Engaged
 - Very Engaged
4. How hard was it to concentrate while you searched?
 - Very Hard
 - Hard
 - Neutral
 - Easy
 - Very Easy
5. Did you encounter any issues while completing this task? If yes, please describe.

If you need to take a break, please do so now. When ready, please click the submit button to continue.

Figure 2: Post-task questionnaire.

| Post-Task Question | p-values of Experiment Factors | | | | |
|--|--------------------------------|-------|-----------|-------|-------|
| | User | Topic | Precision | Phase | Task |
| Finding Relevant Docs (Difficult - Easy) | 0.000 | 0.000 | 0.006 | 0.893 | 0.444 |
| Experience (Unenjoyable - Enjoyable) | 0.000 | 0.049 | 0.016 | 0.003 | 0.345 |
| Mood (Bored - Engaged) | 0.000 | 0.047 | 0.341 | 0.005 | 0.383 |
| Ability to Concentrate | 0.000 | 0.025 | 0.079 | 0.630 | 0.260 |

Table 2: Analysis of variance results for all factors.

| Post-Task Question | P@N = 0.3 | P@N = 0.6 | p-value |
|--|-------------|-------------|---------|
| Finding Relevant Docs (Difficult - Easy) | 2.84 ± 0.08 | 3.13 ± 0.08 | 0.006 |
| Experience (Unenjoyable - Enjoyable) | 2.97 ± 0.06 | 3.16 ± 0.07 | 0.016 |
| Mood (Bored - Engaged) | 3.09 ± 0.07 | 3.17 ± 0.07 | 0.341 |
| Ability to Concentrate | 3.21 ± 0.06 | 3.34 ± 0.07 | 0.079 |

Table 3: Average and standard error of users’ responses given the precision of the results.

3. How would you rate your mood while you searched?
4. How hard was it to concentrate while you searched?

Figure 2 shows the user interface for the post-task questionnaire. In our analysis, we mapped the 5 point Likert scale for each question to the values 1 through 5 with the most negative response mapped to 1, e.g. “Very Difficult”, and the most positive response mapped to 5, e.g. “Very Easy”, and the neutral response mapped to 3.

2.1 Experiment Factors

The factors of our experiment include the phase of the study, the task order, the topic, the user, and the precision of the ranked list. The responses that we examine are the 4 post-task questions that we asked of each user.

Phase The experiment had two phases, phase 1 and 2, as described above. The phase of the experiment contains a possible order effect. Phase 1 of the experiment always occurred before phase 2. The user interface was tied to the phase of the experiment.

Task Users completed two 1 hour sessions. A session corresponded to a phase of the experiment, and each session included 4 tasks. We number the tasks 1 through 4. Search topics and precision levels were rotated across tasks and balanced across all other factors. Of the eight topics, users would complete 4 of the 8 in phase 1 and the remaining 4 in phase 2.

Topic As described above, the experiment used the 8 topics shown in Table 1.

User We continued to recruit participants until after data cleaning [5] we had a completely balanced experiment with 48 users. Each user completed both phases.

Precision We looked at two levels of precision. For a given topic, users would receive either the higher or lower precision ranked list of documents. We carefully constructed the ranked lists of documents to produce near uniform levels of precision at rank N . As such, the ranked lists have near equal precision at N , mean average precision (MAP), and R-precision (precision at the number of known relevant documents, R). The lower level of precision was 0.3 and the higher level of precision was 0.6. We choose these levels of precision based on the range of precision at 10 for the runs without obvious errors submitted to the TREC 2005 Robust track.

3. RESULTS AND DISCUSSION

For each of the post-task questions, we performed an analysis of variance (AOV) with the question as the response. As can be seen in Table 2, precision has a statistically significant impact on perceived difficulty of finding relevant documents ($p = 0.006$). Table 3 reports the average response and AOV p-value for each question given the two levels of precision. As expected, the user and topic also have a significant impact on perceived difficulty and all other responses. The user interface (phase) and task order did not impact perceived difficulty. The task order had no significant effect on any of the responses.

Precision also had a statistically significant effect on the enjoyability of the search experience ($p = 0.049$). It appears that users felt that their concentration was better with higher precision ($p = 0.079$), but this effect was not as strong as for difficulty or enjoyability. There is no evidence that precision affected the mood (engagement) of the users ($p = 0.341$).

The phase did have a significant effect on both the enjoyability ($p = 0.003$) and mood ($p = 0.005$) of the users. Users enjoyed and felt more engaged with phase 2 than with phase 1, but because the user interface was not rotated across phases, we cannot draw any conclusion about the effect of the user interface on the user. We hypothesize that the web-like interface of phase 2 was the cause of the improved enjoyability and mood. There is no evidence of either phase making the user feel as though it was easier to find relevant documents ($p = 0.893$).

To give some sense of the variability in the topics, Figure 3 shows the user performance for both phases and topics. Of the many possible ways to define and measure topic difficulty, one objective measure is the number of documents found relevant by the user in 10 minutes.

Of note, for the results shown in Figure 3, precision is a controlled variable. With an actual retrieval system, users might be able to easily obtain high precision results for one topic but not for the other. For example, Bailey et al. [3] found topics 336 and 367 to be “easy” topics based on the nDCG scores for queries obtained from users. Topic 336 is certainly not “easy” if we take the number of relevant documents found in 10 minutes as our measure of topic difficulty. For phase 2, topic 336 was one of the more difficult topics as measured by number of documents saved as relevant. Topic

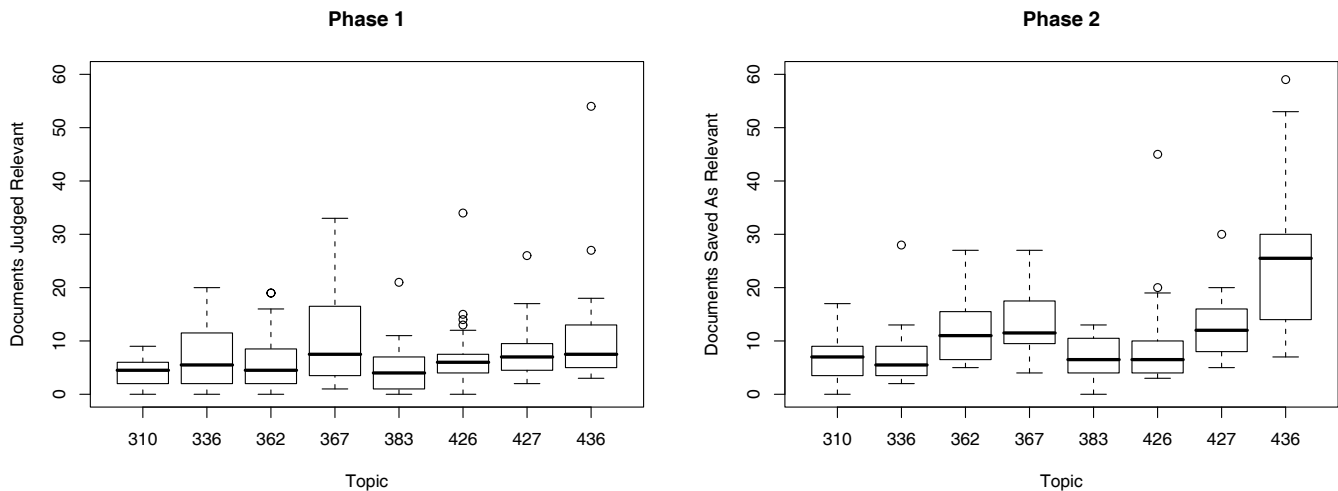


Figure 3: Number of documents judged as relevant (phase 1) and saved as relevant (phase 2) per topic. Each topic represents 24 users’ data. The median is the heavy line inside the box. The box represents the data from the 1st quartile to the 3rd quartile, i.e. 50% of the values fall within the box.

367 does appear to be one of the easier topics. All the topics shown in Figure 3 have an equal number of users searching the same result lists of controlled levels of precision. Our control of precision across topics is one of the differences between our and Bailey et al.’s work.

4. RELATED WORK

There has been quite a bit of research looking at retrieval precision and its effect on user performance and satisfaction. We limit our review here to a few papers on precision’s effect on perceived difficulty and satisfaction. We have already discussed the work of Bailey et al. [3] whose questions we used for our work.

Most similar to our user study is the work of Kelly, Fu, and Shah [4] who controlled the quality of the search results and asked users to evaluate the quality of the retrieval system. Kelly et al. found that higher precision resulted in better evaluation scores for the system. Their study differed from ours in that while we were primarily concerned with user performance, they were concerned with users’ evaluations of the search engines. It may be that users consider retrieval quality to be the same as their difficulty with finding relevant documents, but we think Bailey et al.’s question about perceived difficulty is directed at the user’s personal assessment of their search and not directed towards the search engine. Unfortunately we only had one question about difficulty. Multiple questions would have helped us target some questions towards the user and some towards the retrieval system. In addition to their own work, Kelly et al. provide an extensive and excellent review of related literature.

Al-Maskari, Sanderson, and Clough have studied the relationship between retrieval quality and user satisfaction [2]. They found a strong correlation between user satisfaction, precision, and cumulated gain, but a weaker correlation with discounted cumulative gain, and little correlation with nDCG. There may be an interesting difference between *satisfaction* and *difficulty with finding relevant documents*. Recall that Bailey et al. [3] found a correlation between nDCG and perceived difficulty. In another experiment, Al-Maskari

and Sanderson [1] again report that system effectiveness, as measured by mean average precision, is positively correlated with user satisfaction as is user effectiveness and user effort.

5. CONCLUSION

We conducted a two phase user study that controlled retrieval precision. Across search topics, user performance differed greatly with some topics being much easier than others to find relevant documents. We found that retrieval precision had a statistically significant effect on perceived difficulty of finding relevant documents. In addition, we found that the higher level of precision produced more enjoyment and somewhat increased concentration, but we found no evidence that precision affected engagement with the search.

6. ACKNOWLEDGMENTS

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by an Amazon Web Services in Education Research Grant, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

7. REFERENCES

- [1] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *JASIST*, 61(5):859–868, 2010.
- [2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR’07*, pages 773–774. 2007.
- [3] E. W. Bailey, D. Kelly, and K. Gyllstrom. Undergraduates’ evaluations of assigned search topics. In *SIGIR’09*, pages 812–813, 2009.
- [4] D. Kelly, X. Fu, and C. Shah. Effects of position and number of relevant documents retrieved on users’ evaluations of system performance. *TOIS*, 28(2):1–29, 2010.
- [5] M. D. Smucker and C. Jethani. Human performance and retrieval precision revisited. In *SIGIR’10*. 2010.