Author/s:
Bailey, P;Moffat, A;Scholer, F;Thomas, P

Title:
User Variability and IR System Evaluation

Date:
2015

Citation:
Bailey, P., Moffat, A., Scholer, F. & Thomas, P. (2015). User Variability and IR System Evaluation. Baeza-Yates, R (Ed.) Lalmas, M (Ed.) Moffat, A (Ed.) Ribeiro-Neto, B (Ed.) Proc. 38th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.625-634. ACM. https://doi.org/10.1145/2766462.2767728.

Persistent Link:
https://hdl.handle.net/11343/58275

User Variability and IR System Evaluation

Peter Bailey
Microsoft,
Australia
pbailey@microsoft.com

Alistair Moffat
The University of Melbourne,
Australia
ammoffat@unimelb.edu.au

Falk Scholer
RMIT University,
Australia
falk.scholer@rmit.edu.au

Paul Thomas
CSIRO,
Australia
paul.thomas@csiro.au

ABSTRACT

Test collection design eliminates sources of user variability to make statistical comparisons among information retrieval (IR) systems more affordable. Does this choice unnecessarily limit generalizability of the outcomes to real usage scenarios? We explore two aspects of user variability with regard to evaluating the relative performance of IR systems, assessing effectiveness in the context of a subset of topics from three TREC collections, with the embodied information needs categorized against three levels of increasing task complexity. First, we explore the impact of widely differing queries that searchers construct for the same information need description. By executing those queries, we demonstrate that query formulation is critical to query effectiveness. The results also show that the range of scores characterizing effectiveness for a single system arising from these queries is comparable or greater than the range of scores arising from variation among systems using only a single query per topic. Second, our experiments reveal that searchers display substantial individual variation in the numbers of documents and queries they anticipate needing to issue, and there are underlying significant differences in these numbers in line with increasing task complexity levels. Our conclusion is that test collection design would be improved by the use of multiple query variations per topic, and could be further improved by the use of metrics which are sensitive to the expected numbers of useful documents.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*performance evaluation*.

Keywords

User behavior, test collections, relevance measures

1. INTRODUCTION AND BACKGROUND

In the Cranfield and TREC paradigm, information retrieval test collections (consisting of a corpus, topics, relevance judgments, and a relevance measure – collectively representing a sample of some population of a real-world information retrieval task) allow *comparative* system performance experiments to be carried out. This approach, sometimes referred to as *batch evaluation*, assesses the *system or algorithmic aspect* of relevance, as defined by Saracevic [26]. Almost all sources of variability are removed in this classical design of test collections, including users and tasks, leaving topics as the primary source of variability within the collection. The relevance measure encodes (either explicitly or implicitly) an abstracted model of user behavior, and rewards systems which deliver relevant material more efficiently or comprehensively according to the model. Statistical assessments of comparative effectiveness as determined by the relevance measure can be used to determine improvements in algorithm design. Statistical power analysis calculations can be used to determine the number of topics needed to quantify the probabilities of making type I and type II errors.

An important aspect of test collection use that has perhaps been under-investigated is the degree to which they have *external validity*. Crudely put, external validity characterizes the extent to which an experiment (typically relative system effectiveness according to the relevance measure in the case of test collections) generalizes to other real world circumstances. Such circumstances might encompass use by different users (stay-at-home parents vs retired intelligence analysts) or over different document sets (a subset of the Web vs a collection of news articles) or for different interaction tasks (factoid question answering vs ad hoc topical information discovery). We propose some potential properties of a test collection that relate to the degree to which they have external validity. These include: *fidelity* – whether the relative effectiveness of systems is consistent when a population of users/topics/documents/task (of which the test collection is a sample) uses the systems; *corpus-generalizability* – whether outcomes from this test collection are consistent across other test collections modeling the same task, as investigated by Robertson and Kanoulas [24]; *task-generalizability* – whether outcomes from this test collection are consistent across other test collections with different tasks (i.e., generalizability across situations); and *user-generalizability* – whether outcomes are consistent in the presence of different user behaviors for the same task and topics (i.e., generalizability across people). We provide brief definitions for some other key concepts in Figure 1.

In this work, we are motivated to improve the *user-generalizability* property of test collections. In particular, we seek to understand how introducing some simple sources of variability in users, namely individual query formulation and expectations of the quantities of relevant information needing to be found, might affect how test collections are constructed, and how batch evaluations are carried

SIGIR '15, August 09–13, 2015, Santiago, Chile

ACM 978-1-4503-3621-5/15/08 … $15.00.
http://dx.doi.org/10.1145/2766462.2767728.

Task	An information seeking activity.
Topic	A description of information content or subject matter required for some task.
Task complexity	A categorized model of cognitive information processing for some task.
T	Expected number of useful documents required to satisfy a task for some topic.
Q	Expected number of queries required to satisfy a task for some topic.
Q02, R03, T04	Our subsets of the topics and judgments from TREC test collections corresponding to Question Answering 2002, Robust 2003, Terabyte 2004 tracks.
AP, ERR, NDCG, RBP p, Q β	Various relevance measures: Average Precision, Expected Reciprocal Rank, Normalized Discounted Cumulative Gain, Rank-Biased Precision, and the Q-Measure.

Figure 1: Definitions of significant terms and abbreviations used.

out. We use the lens of *task complexity* (discussed below) to help assess these issues across a range of information seeking scenarios.

To examine *user-generalizability* in a batch evaluation setting, we pose a series of research questions assessed in the sections below.

RQ1 Does the existence of individual variation in initial query formulation for a single information need alter the evaluation of system performance? (Section 4)

RQ2 Is there significant variation among users of the anticipated effort in terms of the number of documents viewed and queries to be issued, and is there a relationship between a user's anticipated effort and the information task complexity? (Section 5)

RQ3 Does incorporating anticipated effort within adaptive metrics lead to changes in relative system performance assessments? (Section 6)

The overarching issue we consider is: to what extent do measures of system effectiveness depend on (lack of) variation in user behavior and thus do test collections have insufficient *user-generalizability*?

We are not the first to consider this issue. In a 1977 report on the design for an ideal test collection, Spärck Jones and Bates [29] recommend that:

> *The effects on the retrieval of relevant documents of such variations over requests should be counteracted by the use of additional queries specifically designed to exhaust the relevant document set.*

The 1999 TREC Query Track examined sets of queries for topics, and the coordinators Buckley and Walz [8] similarly conclude that:

> *We've reaffirmed the tremendous variation that sometimes gets hidden underneath the averages of a typical IR experiment. Topics are extremely variable; queries dealing with the same topic are extremely variable. . . ; and systems were only somewhat variable.*

In a comprehensive study examining different types and sets of judges as the source of user variability, Voorhees [32] found that the TREC-4 and TREC-6 collections were reasonably stable in relative outcomes for participating systems, both for similar users' judgments and different users' judgments. She also observed that inter-system comparisons required more substantial differences in measure scores than for intra-system comparisons. More recently, Bailey et al. [4] examined consequences of using relevance labels originating from judges of differing task and topic expertise. They found that variation in expertise levels led to consistent differences in relevance outcomes and also to questions about the robustness of relative system performance measures over the TREC Enterprise 2007 test collection. Kazai et al. [16] confirmed that such systematic bias between different kinds of judge may exist.

The project described here encompasses exploration of just two (among many) aspects of user variability, thereby to connect user experiences more closely with batch evaluation outcomes.

2. RELATED WORK

Task complexity In information science, the complexity of a search task has long been recognized as having an important impact on information seeking behavior and use, including for example the type and complexity of information needed, and the number and diversity of sources consulted [31].

Byström and Järvelin [9] proposed a five-level task complexity taxonomy, ranging from automatic information processing tasks (tasks that are completely determinable so that they could in theory be automated) to genuine decision tasks (unexpected, unstructured tasks). This taxonomy was refined into three levels by Bell and Ruthven [7], with the distinction between levels being based primarily on the initial determinability and clarity of the task.

Focusing more directly on task complexity in the context of interactive information retrieval, Wu et al. [35] proposed a hierarchy based on the Cognitive Process Dimension of Krathwohl's Taxonomy of Learning Objectives [17], which is itself a refinement of Bloom's Taxonomy of educational objectives. Through a user study, Wu et al. demonstrated a tendency for participants to spend more time, issue a greater number of queries, and click on more search results for tasks with greater cognitive complexity. We use three levels of this taxonomy for our experiments, explained below.

Factors that influence searcher behavior Wu et al. [36] investigated the relationship between information scent (signals of relevance on a search results page) and search behavior such as query reformulation, search depth and stopping, demonstrating that a higher density of relevant items on the first page increases the probability of query reformulation, and decreases that of pagination.

The relationship between constraints and searcher behavior was studied by Fujikawa et al. [13], who showed that when the number of queries that a searcher can enter is restricted, greater attention is given to query formulation and more time is invested in viewing search results pages. Similar effects were observed when constraints were placed on the number of documents that can be viewed.

Azzopardi et al. [3] studied the effect of query cost on the behavior of searchers, examining the influence of different interfaces, designed to require differing amounts of effort. Users of the "structured" (highest cost) interface displayed different behavior, submitting fewer queries and spending longer when examining search result pages. A strong relationship between searcher behavior and task type and structure was also reported by Toms et al. [30], with users showing different rates of query reformulation and page views.

In a focused study, White and Kelly [34] varied the threshold acquired from individual document examination times as an input to an implicit relevance feedback algorithm, across a number of individuals and search tasks. They found that there was substantial variation in individual examination times, and that it was possible to improve relevance performance by using task information to determine the threshold. Attempts to tailor the threshold on a per-individual basis led to degraded performance however, suggesting intra-task-consistency was higher than intra-individual-consistency.

Gwizdka and Spence [14] examined observable measures of information seeking activities (such as documents viewed, time spent etc.) of a set of psychology students within a laboratory setting. They characterized relationships between the objective operationalized task complexity (in a manner influenced by [7]) and subjective searcher assessments of task difficulty with respect to these observable measures, and analyzed which measures were more important in predicting the difficulty experienced by the searcher. They found that task complexity affected both the relative importance of these predictors and the subjective assessment of difficulty. They also observed that individual variation (in factors like experience, verbal ability, other cognitive abilities etc.) played an important part in affecting performance and relative assessment of difficulty. We use individual variation in query formulation and expected goals of search to examine how batch evaluation outcomes change, and use task complexity as an analysis factor.

Query variability Searchers use an IR system to resolve an information need. To do so they need to translate their internal information requirement into an explicit query that is submitted to the search system. Multiple queries can represent a single information need, and indeed a single user may issue multiple queries within a single search session. Finally, interactive query (re-)formulation systems are increasingly common and have been demonstrated to assist in improving retrieval performance by (among others) Kumaran and Allan [18]. In that work, the authors also demonstrate how programmatic query expansion or relaxation can lead to significant increases in performance, across a selection of TREC test collections.

The 1999 TREC Query Track [8] investigated the issue of query variability through the creation of 23 query "sets", alternative query statements corresponding to 50 TREC topics. Analysis confirmed previous research showing that differences between topics introduces substantial variability into IR experimental results, and further showed that the variability of queries dealing with the same topic also introduced significant variability, typically greater than differences between retrieval systems. However, Buckley and Walz note that formal conclusions cannot be drawn from the full data set, due to the presence of "blundered queries" and the presence of multiple versions of the same basic system [8]. Other investigations of query variability in the TREC setting were shown to improve query performance through data fusion [5, 6].

Modave et al. [20] carried out a study of the quality of health-related information related for people seeking information about weight-loss using Google. While measuring query variability was not a focus of the study, this effect was accounted for by generating a range of queries about the weight-loss topic, eliciting specific queries from 20 study participants as well as the Google auto-complete feature.

Evaluation metrics Batch evaluations rely on objective scoring of search response listings. Long-standing mechanisms include Reciprocal Rank (RR); Precision at depth k; and Average Precision (AP), the average of the precisions achieved at the depths in the ranking of the relevant documents. A wide range of further alternatives have been developed over the last decade, including Normalized Discounted Cumulative Gain (NDCG) [15]; Rank-Biased Precision (RBP) [22]; Expected Reciprocal Rank (ERR) [10]; and the Q-Measure [25]. Per-query scores from one or more of these metrics are then averaged in some way, and paired statistical tests applied in order to draw experimental conclusions.

Metrics are sensitive to system performance in different ways. Precision at 10 and RBP with parameters less than about $p = 0.8$ are "shallow" metrics, and hence better match the behavior of a typical web search user than do "deep" metrics such as AP, NDCG, and the Q-Measure. In terms of judgment effort, shallow metrics are also cheaper to evaluate than deep metrics. On the other hand, deep metrics tend to lead to a higher fraction of statistically significant system differences being identified (the discrimination ratio), and to be just as predictive of the behavior of shallow metrics as are the shallow metrics themselves [33]. Moffat [21] provides further commentary on ways effectiveness metrics can be categorized.

User goals and persistence Users vary in the way they process search response pages, and hence if a metric is to reflect the user's perception of their experience, should be sensitive to that variation. Moffat and Zobel [22] argued that a metric should match a *user model*, a description of the behavior of the presumed user; and parameterized their RBP metric with a persistence parameter p. Rather than quantifying persistence in terms of documents, Smucker and Clarke [28] used time as the primary persistence factor in the model, and make use of data from a user study to calibrate their gain calculations. Moffat et al. [23] note that users may have differing goals, even for the same query or same information need, and introduce the notion of an "expected goal of search", their parameter T, and use it to shape predictions about what happens when that user is viewing a page, thereby creating a more refined user model that in turn leads to further alternative effectiveness metrics. A user study provided evidence to support that hypothesis, bringing user and batch evaluations a step closer. We build on their work by examining how user variation in queries and expected goals can be combined. Next we describe our overall experimental framework.

3. EXPERIMENTAL FRAMEWORK

Search can be viewed as a process that starts with an information need, out of which a particular query is formulated by a user and submitted to a retrieval system. However, batch evaluations typically start with a single query per information need and regard the system as being the primary variable that impacts on effectiveness. Our experimental framework attempts to reintroduce two aspects of user variability into the batch evaluation process. We start by describing the process we adopted for formulating information need statements that could then be used to investigate user-generalizability.

Information needs To investigate user generalizability, several aspects of searcher behavior were studied through a crowd-sourced experiment. We first required a set of labeled search tasks for the experimental participants to carry out. To obtain a broad cross-section of information-seeking tasks, a set of 180 TREC topics was selected:

- **Q02** Question Answering Track 2002, 70 topics (1824–1893)
- **R03** Robust Track 2003, 60 topics (selected from 303–610)[1]
- **T04** Terabyte Track 2004, 50 topics (701–750)

For each topic, a *backstory* was created; this was a short information need statement that was intended to motivate and contextualize the search request, making the topic statements less abstract and more engaging. Four annotators created the backstories, based on the full original TREC title, description and narrative fields. They were also free to explore related background information using online resources. An example topic from each of the three TREC tracks is shown in Figure 2. To encourage our eventual experimental participants to engage more fully with these search tasks, and to treat them as personal searches rather than abstract impersonal ones,

[1]The topic numbers are non-contiguous because half of the topics selected for the Robust Track 2003 were chosen as they were known to be difficult from previous Ad-Hoc tracks.

Q02.1828, Remember*; "What was Thailand's original name?"*

While visiting Thailand for a beach holiday last year, you decided to visit some local museums to learn more about Thailand's history. You learned many interesting things about the country, including that it was not always called Thailand. What was it called originally?

R03.356, Understand*; "postmenopausal estrogen Britain"*

A friend, who lives in Britain, has started estrogen treatment. This surprises you as you thought it's no longer recommended. You want to find out more about the use of hormone replacement therapy or estrogen treatment in the U.K.

T04.734, Analyze*; "Recycling successes"*

Your city has recently embarked on an ambitious zero-waste policy for household and industrial garbage. Recycling is going to be a big component of the program. You wish to find out what recycling projects have been successful, including the places or product programs that have worked, and what they understood success to mean.

Figure 2: Backstory associated with three TREC topics from different tasks in different years, together with the task type.

Number: 734

Recycling successes

Description: What recycling projects have been successful?

Narrative: Guidelines by themselves are not relevant. Titles in a table of contents are relevant if they identify places or product programs which have had success. Must be declared successful or success should be clearly assumed from the description. Name of state identified as successful recycler is relevant. Listing of recycled products for sale are relevant.

Figure 3: Topic 734 from the TREC 2004 Terabyte Track.

the backstories were written to speak directly to the reader, and to include hypothetical family members or friends. Figure 3 shows the original TREC presentation of one of the topics shown in Figure 2.

The original topic statements from the Terabyte and Robust tracks contain substantial detail about what information a document should or should not contain to be considered relevant, and the created backstories aimed to reflect the bulk of these requirements. Nevertheless, we acknowledge that there is potential for drift between the interpretations of the backstory and the original TREC topic description that led to relevance judgments being created. Topics from the QA Track were more difficult as they are typically presented simply as question statements, such as "*How much gravity exists on mars?*" (Q02.1871). Simply posing the question statement to the experimental participants might lead to these being entered directly as a search query, rather than then being read as an information need statement, so the QA topics are also presented with a backstory. When possible, pronouns or other indirect references to the query subject were used, to reduce the likelihood that participants would simply copy and paste the final question as their query.

Task complexity Different information-seeking tasks have different characteristics, and task complexity is a key feature that may influence searcher behavior. For our experiments we adapt three levels from the cognitive complexity hierarchy proposed by Wu et al. [35], derived from a taxonomy of learning objectives presented by Anderson and Krathwohl [2]. This hierarchy considers a spectrum of information needs, with the lowest level consisting of searches that involve *"retrieving, recognizing, and recalling relevant knowledge"*. Such *Remember* queries therefore involve finding a fact in response to a simple "when", "where" or "what" question, such as "*How did Eva Peron die?*". The next level in the hierarchy, *Understand*, involves *"constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining"*. We also use a third level, *Analyze*; tasks at this level of the hierarchy involve *"breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing"*. Each of the example tasks in Figure 2 indicates the corresponding complexity category.

Based on the created backstories, each of the 180 selected TREC topics were assigned to one of the task complexity types. Four annotators independently rated each topic: broadly speaking, topics that required a simple factoid answer tended to be assigned to the *Remember* category; where topics required the production of list of things, even if relatively complex and sourced from different pages, they tended to be assigned as *Understand*; and, where topics required synthesis of disparate information, and eventual summary, or balancing of competing viewpoints and opinions, they were allocated to the *Analyze* category.[2] The overall inter-annotator agreement among the four judges for the initial ratings was 0.664, measured by Fleiss' κ [12], a statistic that measures agreement across multiple raters and corrects for agreement by chance. It is interesting to note that the per-category agreement varied substantially, from 0.456 for the highest of the three hierarchy levels (*Analyze*) to 0.907 for the lowest level (*Remember*), indicating that *Remember* tasks are relatively easy to identify and agree on, while differentiating between *Remember* and *Analyze* tasks is more difficult. For all cases where there was no majority rating among the four annotators, the tasks were carefully discussed until agreement was reached, resulting in a single confirmed type for each.

Gathering data To investigate user variability in a test collection setting, an experiment was carried out using the CrowdFlower crowd-sourcing platform.[3] The experiment was reviewed and approved by the <anonymous institution> ethics board.

On signing up for the experiment, a participant was first presented with an information need statement, one of the created backstories. They were then required to answer three questions. First, participants were asked: *How many useful web pages do you think you would need to complete the search task?*. Responses were selected from the following: 0 useful pages (I'd expect to find the answer in the search results listing, without reading any of the pages); 1 useful page (I'd expect to find the answer in the first useful page I found); 2 useful pages; 3-5 useful pages; 6-10 useful pages; 11-100 useful pages; 101+ useful pages. Second, they were asked: *In total, how many different queries do you think you would need to enter to find that many useful pages?* with answers selectable from the following: 1 query (I'd expect to be able to complete the search task after the first query); 2 queries; 3-5 queries; 6-10 queries; 11+ queries. Third, participants were asked: *What would your first query be?*; answers to this question were entered in a textbox. Participants were free to complete as many topics as they liked, from one to a maximum of 180. The resulting data set had 10,800 responses from 115 workers,

Cleaning crowd data It cannot be expected that all anonymous

[2]To promote reproducibility, the full set of 180 topic backstories and corresponding task complexity labels will be available on request.

[3]`http://www.crowdflower.com`

	Task complexity		
	Remem.	Under.	Analyze
number of topics	70	81	29
average queries per topic	44.3	44.4	44.0
average query length (chars)	25.9	32.9	37.1
average query length (words)	5.6	6.0	6.5
average query entropy (bits)	19.9	26.0	30.5

Table 1: Query properties after normalization: average query length in characters, not counting white-space characters; average query length in words; and average query entropy in bits. To calculate the last, the frequency distribution of words appearing in the queries for each topic was computed, and then the average information cost of representing the queries for that topic computed using that frequency distribution, and averaged over task complexities.

workers took their task seriously, and where it was possible to identify clearly inappropriate responses, those workers were removed from further analysis (but still paid). First, if any worker suggested the same first query for two or more tasks, they were considered unreliable and all their responses were removed – recall that no worker got the same task twice, so it is extremely unlikely that two tasks would attract identical queries. This rule removed 15 of 115 workers. Two further workers who had copy/pasted apparently nonsensical parts of the topic statement as their first query were also identified and removed. This left 7,971 responses from 98 workers, covering all 180 topics with 41–48 responses per topic (median 44).

4. VARIATION IN FIRST QUERIES

Having described the data collection process, we first examine the sets of queries suggested by the experimental subjects.

Normalization One of the components in each interaction pane asked "*What would your first query be?*" Workers then entered text in to a textbox. As with all web queries, the resultant strings are noisy, with a wide range of spelling and grammatical errors. In this regard, the behavior of the crowd workers probably corresponds closely to other users. To ameliorate this type of behavior, web search systems include a "did you mean?" query modification feature. To faithfully reflect that behavior, the query strings typed by the crowd-sourced subjects were converted to US English, and corrections applied whenever they could be unambiguously identified. For example, "theropy" was changed to "therapy" in the context of topic R03.356 (Figure 2). In some cases the correction was not clearcut, or the erroneous word was actually a correct spelling of something different. Manual interactions with a major search engine were used to decide whether to alter these queries. For example, "cheapskate bay" was altered to "chesapeake bay", because that is what happened at a web search interface. On the other hand, "calgary provicence" was altered to "calgary providence" rather than "calgary province", which would have better fitted the topic in question, because the first alteration was what was suggested at the same search interface. As a further part of the normalization process all punctuation characters were removed, including periods. Finally, two queries ("zdvfdzfvg" and "fxghfsdg") not caught by the earlier quality-control mechanisms were removed. The resulting query set contained 7,969 queries, of which 5,046 were unique.

Query diversity Table 1 lists some properties of the queries received, averaged over the three query classes after quality control and normalization mechanisms were applied. The table shows a clear trend to longer queries as the information need becomes more complex, both in terms of characters typed and in terms of words typed. The final row of the table represents the average diversity of the terms across the pool of queries generated for each topic, by computing the term frequencies of all terms used in queries for that topic, then calculating the entropy of each query relative to that distribution, and finally averaging those average entropies. The entropy of a query increases as the length of the query increases, and is also high if a broad set of term is being used across the pool of queries for that topic – if queries are less predictable. This measure confirms that the more complex the information need, the more expressive are the queries posed to resolve it.

city recycling projects (2)
city recycling scheme progress
council website
most successful recycling programs
recycling policy update
recycling projects (2)
recycling projects for household and industrial garbage
recycling projects program
recycling projects successes and effects
recycling projects that have been successful
recycling successes
reducing waste to zero success stories
successful city recycling policies
successful municipal recycling projects
successful recycling programs (2)
successful recycling projects (11)
successful recycling projects place product programs
successful zero waste
what are the recycling projects that have been successful
what does it take to make a successful recycling program
what recycling projects have been successful (6)
where have recycling projects been successful and how do they define success
zero waste policy (2)
zero waste policy for household and industrial garbage
zero waste policy for household and industrial garbage programs

Figure 4: The 44 user-generated queries for Topic 734 (Figure 3). Numbers in parentheses indicate multiplicity.

As a single example, Figure 4 lists the complete set of queries generated for one of the 180 information need statements (see Figure 2). One query dominates – an extended version of the TREC title-only query for this topic, "recycling successes" – but nearly half of the queries generated by the subjects occur only once.

Query effectiveness – In the small Two different retrieval systems were then used to execute each query against the corresponding document collection: Indri[4] with an Okapi similarity computation, and Indri with a sequential dependency computation [19]. Using Indri for both ranking algorithms ensures the system effects are due to fundamental differences in the retrieval algorithms, rather than other factors related to query or document processing. Rankings of length 200 documents were generated and scored; with documents for which no judgment was available deemed to be not relevant.

Figure 5 shows the range of scores that resulted when four standard relevance measures were applied to the rankings for the set of 44 queries generated in response to Topic 734 (Figures 2 and 4), with the Indri Okapi BM25 and SDM ranking functions. The blue diamonds show the corresponding scores for the canonical TREC

[4] `http://www.lemurproject.org/indri/`.

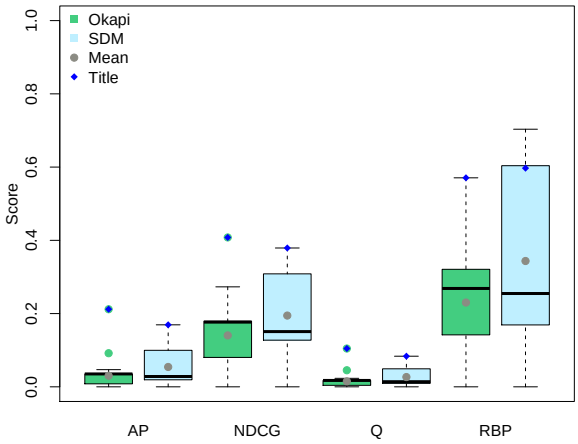

Figure 5: Retrieval effectiveness measured by AP, NDCG, Q1 and RBP0.85 for Topic 734. Green and blue boxes show scores obtained from running the different user queries using BM25 and SDM, respectively. Grey points indicate the mean for each column, black bars the median, and the blue diamonds show the effectiveness of the corresponding TREC title-only query.

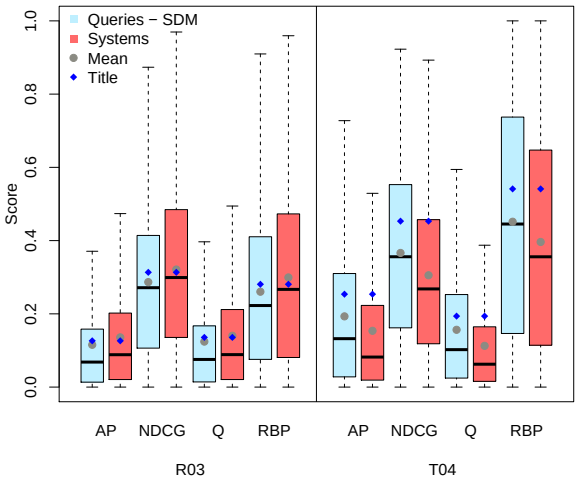

Figure 6: Retrieval effectiveness as measured by AP, NDCG, Q1 and RBP0.85, for two subcollections R03 and T04. Blue boxes show scores obtained from running different user queries with SDM retrieval, while red boxes show scores achieved by different TREC contributing system runs. Grey points indicate the mean for each column, black bars the median, and the blue diamonds show the effectiveness of an Indri SDM run using the corresponding TREC title-only query. The average residuals for the four RBP measurements were (left to right) 0.153, 0.038, 0.235, and 0.059. The Indri SDM runs had RBP residuals of 0.017 for R03, and 0.056 for T04.

title-only query (again, see Figure 2) when evaluated using the same two retrieval mechanisms. The results for this one query reflect a trend that we also saw more widely – that for typical user queries the mean performance of SDM is superior to that of Okapi. That difference is consistent, but not absolute, and for most combinations of metric and topic there are also queries for which Okapi out-scored SDM. On Topic 734 the title-only query was the highest-scoring query (of the 44) for AP, NDCG, and Q1 for both Okapi and SDM models, and also the highest-scoring for RBP0.85 for Okapi (the query "successful recycling projects place product programs" scored 0.703 when the SDM similarity model is used), but this topic was unusual in that regard. For example, for Topic 356 (Figure 2), more than half of the user-generated queries outperformed the canonical title-only query. The omission from the corresponding backstory of the word "postmenopausal", which appears in the TREC topic description ("identify documents discussing the use of estrogen by postmenopausal women in Britain"), may have had an effect. Some level of unintentional topic drift is always possible in our process.

A risk factor in any experimentation in which judgments are reused is the extent to which they provide coverage of the documents retrieved by the systems being compared. For Topic 736, the RBP *residuals* when $p = 0.85$ are 0.518 and 0.469 for Okapi and SDM, respectively. These represent the assessment weight of the unjudged documents in RBP [22], with 0.0 representing a situation with all required judgments available, and larger values indicating that the RBP score would increase by that much if all unjudged documents were in fact relevant. For Topic 734 the average residual for the user-generated queries was in excess of 0.25, and the available judgments covered less than 75% of the RBP probability mass. That is, the relativities shown in Figure 5 need to be taken cautiously. With more complete judgment coverage, the RBP scores for this topic (and hence the scores for other metrics) might change considerably. Similar situations were encountered for several other topics. On the other hand, the title-only queries have consistently low residuals, because they were used by some of the systems that contributed to the pools from which the judgments were created.

Query effectiveness – In the large The Q02 queries were especially prone to the problems arising from sparse judgments, with RBP0.85 residuals that averaged around 0.5 over that set of 70 topics. One reason may be that QA relevance was focused on answer fragments, not document-level relevance. Coverage was somewhat better on the R03 and T04 queries (but with considerable variation, as already noted).

Figure 6 incorporates score information derived from all of the TREC participating systems that contributed runs for the Robust-03 and Terabyte-04 tracks. The boxes show variation over the corresponding R03 and T04 topics of score responses to user-generated queries (blue) and score responses across the set of contributed TREC runs for that year (red), factoring in all of the user-generated queries when evaluated using the Indri SDM model. The average of the Indri scores for the title-only queries for those topics is also marked on each bar. It is clear that query-derived variations are just as broad as are the variations caused by system diversity, and hence that improved performance relative to the Indri SDM title-only runs is thus equally likely to be derived from query reformulation as it is from system improvement. Note also that for the user-generated queries (blue boxes) there is a considerable amount of metric weight still sitting in the residuals, which might be released with further judgments, and result in higher scores.

Variability analyzed The effect of query choice is illustrated further in an analysis of variance for each metric, modeling score as a response to topic, system, and query. In this analysis "topic" is a nominal variable, one level per TREC topic; "system" has one level for each TREC system, plus two levels for our Indri runs; and "query" has one level for all TREC systems plus one for each query processed by Indri. (We do not know the exact query used by each TREC system, but by assuming it is always the same we will underestimate the variability due to query phrasing and overestimate that due to system.) Q02 looks very different, as discussed above, and those runs are not included in this analysis since these measures are document-relevance centric.

Metric		η^2	SS	df	F
AP	query	0.55	158.40	4977	4.58
	system	0.20	32.61	147	31.90
	topic	0.14	22.05	179	17.72
NDCG	query	0.59	279.02	4977	5.38
	system	0.28	75.64	147	49.35
	topic	0.16	36.56	179	19.59
Q1	query	0.57	145.79	4977	5.05
	system	0.19	24.83	147	29.12
	topic	0.18	23.63	179	22.76
RBP0.85	query	0.53	341.80	4977	4.23
	system	0.20	77.12	147	32.33
	topic	0.11	38.90	179	13.39

Table 2: ANOVA for four metrics, modeled as a response to system, topic, and query string. Partial η^2 values reported; all F statistics significant at $p \ll .001$. In each case, the effect due to query phrasing is substantially larger than that due to topic or system.

Table 2 summarizes the results. Each of query, system, and topic has a statistically significant effect ($p \ll .001$ in all cases, Wald test) and the effect of each factor is medium/large, with the possible exception of topic for RBP, but the effects are of very different scales. The variation due to system is slightly larger than that due to topic (e.g. partial η^2 of 0.20 and 0.14 for AP), so slightly more variation in final score is explained by changes to system than by changes to topic. The variation due to query phrasing, however, dwarfs other effects and over 50% of variation in final score can be attributed to phrasing *even after system and topic are taken into account* (partial η^2 in the range 0.53–0.59).

Observations Particular choices of query clearly can lead to widely different scores, independent of the topic, the system, or the metric. We commonly want to use variation in scores to say something about differences between systems (i.e., "system B is better"); less commonly, we want to use variation in scores to say something about topics (i.e., "topic 734 is hard"). In either case we need to be aware of query wording as a confound, and an extra source of variation which in fact dominates system and topic.

Two macro implications can be drawn from this analysis regarding test collection design and development. First, since this approach supplies between one and two orders of magnitude more queries for a given set of topics within a collection, even given some cross-query document overlap when judging pools on a per topic basis, this would sharply increase the required judging budgets. Given finite budgets, this implies that measures that accommodate missing judgments such as RBP or the suite of inferred AP and NDCG measures developed by Yilmaz and Aslam [37] are required and/or more cost-effective judgment acquisition methods such as crowd-sourcing approaches (e.g., as discussed by Alonso et al. [1]) should be employed. Second, systems could be provided with each topic's collection of queries, and can then make use of any methods desired to create a single top-K ranking for the topic. Document pools would be formed in the usual way, but on the scale of number of topics, not number of queries. In the absence of search engine logs, this might provide some partial subset of the data that is available to commercial search providers about variant phrasing, and hence techniques such as pseudo relevance feedback or query reformulation merging [27] could be employed.

Factor	Effect (mult. odds) T	Q
Worker	0.005–7520.3	10^{-9}–22316.9
Author	0.8–1.3	0.9–1.5
Remember (baseline)	1.0	1.0
Understand	14.1	11.2
Analyze	21.9	18.9

Table 3: Significant factors in fitted models for estimates of T and Q. Effect sizes > 1 correspond to higher values of T or Q being more likely. All effects significant at $p < 0.05$, Wald test.

5. VARIATION IN EXPECTATIONS

As well as differing in their behavior (issuing different queries, in our example), users may have different expectations of a search system and of a task and it would be appropriate to consider this when evaluating search systems. For example, if one user expects to issue a query then read three or four documents – perhaps to compare information from different sources – then it would not be appropriate to evaluate based on the rank of only the first relevant result. If another expects to issue several queries in succession, then it may be appropriate to evaluate a session rather than a single question. Other, similar, scenarios are easy to imagine. Two questions in our instrument aimed to understand some of these varied expectations.

Expected number of documents Cooper [11] noted that (p.31) "most measures do not take into account a crucial variable: the amount of material relevant ... which the user actually needs". Following Moffat et al. [23], we denote this quantity by T. Cooper further observes (p.33):

> *A search request is therefore to be conceived in the abstract as involving two parts: a relevance description (normally a subject specification) and a quantity specification. To put it another way, every search request has a definite quantification.*

To understand this quantification and how it varies, we asked: "how many useful web pages do you think you would need to complete the search task?". We plot the responses for each task complexity category in Figure 7. The distribution of responses across the three types of task are significantly different ($\chi^2 = 2067.0, df = 12, p \ll .001$), and it seems that descriptions of more complex tasks prompt people to expect more reading.

Clearly estimates of T vary with task complexity, and they may vary with other factors as well. Some of these factors are captured in our instrument, and some are external and not captured. To clarify some of the instrument-captured factors, we used cumulative logistic regression (also called ordinal regression) to model T as a response to a number of potential explanatory variables: complexity (three levels), worker (98 levels, or one per worker), topic author (four levels), and CrowdFlower run (two levels). Model selection was performed to minimize the Akaike information criterion (AIC), which measures likelihood but with a penalty for complex models.[5]

The model is summarized in Table 3, where effects are given as multipliers to odds ratios. Effects greater than 1 represent a higher probability of answers higher up the scale. For example, a multiplier of 2 would mean the odds of estimating T as 1 vs estimating T as 0 are twice as high as the baseline; likewise, the odds of estimating T as 2 vs either 1 or 0 would be twice the baseline, and so on.

The largest effects are due to the CrowdFlower workers. Our workers varied substantially, with a worker at one extreme claiming

[5]Modeling used R's `ordinal::clm` and `ordinal::step.clm` functions.

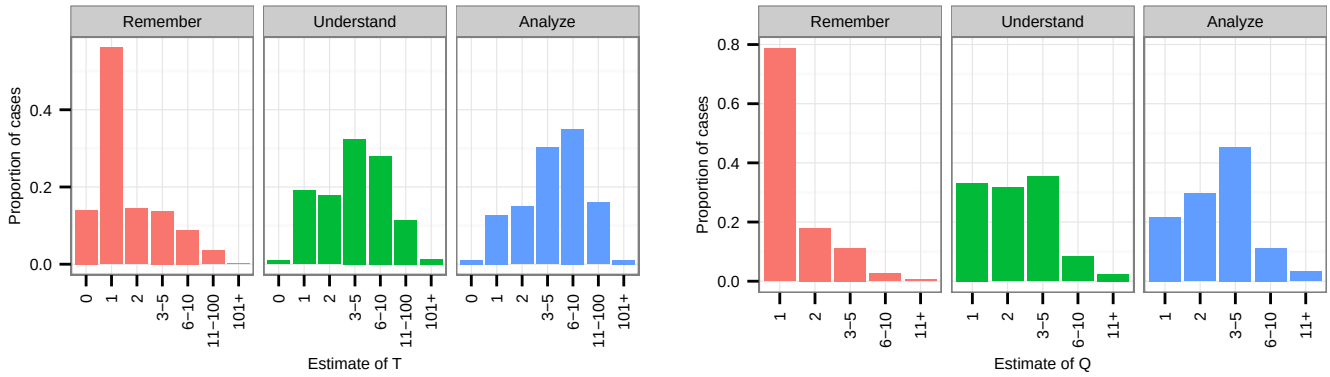

Figure 7: Judges' estimates of T, the number of relevant documents they expect to read, vary with task complexity (left); as do their estimates of Q, the number of queries they expect to issue (right). Judges expect to need more interactions for more complex tasks.

they expected to read documents for only five out of 65 topics and one worker at the other extreme expecting to read 11 or more documents in every case. This is reflected in the model, where per-worker effects are highly variable and again dominate all other effects – odds ratios change by six orders of magnitude for T.

There is a smaller but still notable effect due to task complexity, with *Understand* tasks more likely to prompt higher estimates of T and *Analyze* more likely still – the difference between *Remember* and *Understand* being larger than that between *Understand* and *Analyze*. Finally, there is an effect due to topic author: even after controlling for task complexity and worker, some authors provoked higher T estimates, an effect that was statistically significant, but practically negligible. We also checked for batch effects (the tasks were released to workers in two rounds), but they were not evident.

Expected number of queries Users may also vary in the number of queries they expect to issue – be it a single query, if they think a task is simple or well-supported, or very many in succession, if they think the task is more complex. We denote this expected number of queries Q and plot it in the right-hand side of Figure 7. As before, per-judge effects dominate – the variability here is even more pronounced than for T, with odds ratios varying across thirteen orders of magnitude. Clearly different users have very different expectations of their search engine interactions, even for the same topic. There are again significant differences across complexity levels, with similar effects to those seen for T. Again the difference between a *Remember* and an *Understand* task is larger than the difference between *Understand* and *Analyze*. We also note a significant but small effect due to topic author, and no significant batch effect.

Observations Just as with individual variation in query formation, we observed significant individual variation in expectations of documents and queries to satisfy that need. Given that T is correlated with task complexity and is strongly influenced by user-specific expectations, a natural question to ask is whether it is possible to include T – and progress made towards attaining T as a ranked list is consumed – into an relevance measure in a meaningful way. In the next section, we develop this possibility using the data we collected.

6. SENSITIVE EVALUATION

Current effectiveness metrics are insensitive to both T, the initial expectation of the user, and to the evolving expectation of T as the search is prosecuted. For example, while the p parameter of RBP [22] provides adjustment for user persistence and can be adjusted so as to influence the expected depth in the ranking that the user will examine, the user model associated with RBP requires that the user proceeds to the next document with fixed probability, regardless of how much information has already been accumulated, or what depth has been reached. We believe that as more relevance is accumulated, the user becomes less likely to continue their search.

Expected search length Cooper's definition of ESL [11] is simple: it is the total number of documents inspected before T relevant ones have been found. As such, it is always greater than or equal to T, with larger values representing inferior performance. To obtain a metric with the usual behavior (bounded by zero and one, with larger values indicating better performance) we scale by T and invert, to obtain an ESL- and RR-inspired metric:

$$\mathrm{RRT}(T) = \frac{T}{\text{rank of } T \text{ th relevant document}}. \qquad (1)$$

As is the case with RBP, Prec, and RR, the score returned by this metric corresponds to the average rate at which utility (relevance) is acquired per document inspected, with a user model defined by a person who seeks exactly T relevant documents, and stops their search immediately upon finding a T th answer in the ranking. Queries that have fewer than T answers in a system's ranking are scored as zero. There is a clear connection between RRT and precision – RRT is equal to precision at the T th relevant document. Note also that $\mathrm{RRT}(1)$ is exactly reciprocal rank, RR.

Probabilistic users A second option is to form a probabilistic composite of RBP and ERR [10]. Suppose that the user makes a biased decision after encountering each relevant document, continuing to scan with probability $p = (T-1)/T$, and ending their search with probability $(1-p) = 1/T$. In this user model the expected utility per document inspected is given by:

$$\mathrm{ERRT}(T) = \frac{1}{T}\sum_{t=1}^{\infty}\left[\left(\frac{T-1}{T}\right)^{t-1}\cdot \mathrm{RRT}(t)\right]. \qquad (2)$$

The geometric distribution means that the average number of relevant documents identified by the time the user stops scanning is $1/(1-p) = T$, and that the value of ERRT is non-zero even if fewer than T relevant documents appear in the run. There is also a clear relationship between ERRT and AP, the latter being an unweighted sum of all R precision scores: $\mathrm{AP} = (1/R)\sum_{t=1}^{R}\mathrm{RRT}(t)$, where R is the number of relevant documents for that topic. A key difference between AP and ERRT is that computation of the latter does not require knowledge of R. Nor is AP sensitive to T, of course.

Moffat et al. [23] have also considered effectiveness metrics that are sensitive to T. Their INSQ and INSQ′ functions are weighted precision metrics defined in terms of the conditional probability $C(i)$

T	Upper	Lower		
		INSQ	INSQ′	INST
1	2.58	2.58	1.64	1.33
3	6.53	6.53	4.36	3.27
10	20.51	20.51	13.93	10.26
30	60.50	60.50	41.29	30.25

Table 4: Expected search length for INSQ-based metrics for different values of T, when no documents in the ranking are relevant (column "upper"); and when every document is (columns "lower").

of the user continuing from the document at depth i in the ranking to the document at depth $i+1$. They define INSQ and INSQ′ via

$$C(i) = \left(\frac{i+2T-1}{i+2T}\right)^2 \text{ and } C'(i) = \left(\frac{i+T+T_i-1}{i+T+T_i}\right)^2,$$

respectively, where T_i is the amount of relevance (or gain) that has not yet been accumulated by depth i, $T_i = T - \sum_{j=1}^{i} r_i$, and where $0 \le r_i \le 1$ is the relevance of the ith document in the ranking. Both versions of INSQ are sensitive, in that higher values of T lead to more patient search behavior and a greater expected depth in the ranking. In addition, INSQ′ is *adaptive* – as relevant documents are identified, the expected remaining search cost decreases.

INST: An improved INSQ′ In the formulation of Moffat et al. [23], T_i is required to be positive. We remove that restriction, and allow T_i to be negative too, covering situations in which more gain has been accrued than was initially anticipated. The altered metric, still using the continuation function $C'(i)$, is denoted as INST. Table 4 shows why we prefer this change: it lists expected search depth for INSQ, INSQ′ and INST in two extreme situations – when no documents in the ranking are relevant, and when every document in the ranking is relevant. As is evident in the table, INSQ is not adaptive, and has the same behavior in both extreme situations, examining an average of $2T+0.5$ documents. On the other hand, the expected search length in INSQ′ and INST decreases if relevant documents are encountered. Our preference for INST over INSQ′ is based on its expected search length of approximately $T+0.25$ when all documents are relevant, intuitively a better fit than the approximately $1.4T$ expectation of INSQ′. That is, INST anticipates that a user seeking T relevant documents will examine, on average, between T and $2T$ documents before leaving the ranking, with the actual exit depth depending on the number (and locations) of the relevant documents. In addition, INST retains the other features that made INSQ a more representative model for user behavior [23]. Compared to the INSQ/INST variants, RRT and ERRT give rise to models in which the user may only exit the ranking as they encounter each relevant document. On rankings that do not contain any relevant documents at all, the models associated with RRT and ERRT (like RR and AP before them) have the user scanning the full collection.

Retrieval effectiveness We use RRT(T), ERRT(T), INSQ(T), and INST(T) in our experimentation. To set the parameter T, the distribution of T-bands indicated by the crowd-workers for each topic is employed, and the mapping $0 \to 1$, $1 \to 1$, $2 \to 2$, 3–5 $\to 3$, 6–10 $\to 6$, 11–100 $\to 11$, and 101+ $\to 101$. That is, each score computed for RRT, ERRT, INSQ, and INST is a weighted average of up to six different T-based scores.

Table 5 lists scores for the 110 R03 and T04 queries topics and the user-generated queries, as measured by the four T-sensitive metrics; together with the expected depth at which users exit the result ranking. All of these metrics allow residuals to be computed;

Metric	Score	Residual	Depth
RRT	0.421±0.267	0.195±0.188	48.98
ERRT	0.453±0.267	0.114±0.113	44.17
INSQ	0.310±0.213	0.251±0.185	10.46
INST	0.366±0.249	0.206±0.186	8.57

Table 5: Averages and standard deviations of topic means for 4,871 user-generated queries over 110 R03 and T04 topics, using Indri SDM retrieval, and weighted distributions of T for each query. The final column lists the expected retrieval depth in the corresponding user models, also as a weighted average.

the variation in queries means that these are again relatively high compared to title-only queries. Note that the residuals associated with INST are smaller than those of INSQ, a consequence of the shallower expected depth; and the relatively implausible evaluation depths associated with RRT(T) and ERRT(T) (these would be even higher if the rankings were extended beyond 200 documents).

Kendall's τ for TREC systems Table 6 lists Kendall's τ-b coefficients, computed by scoring TREC systems using a total of nine metrics, and ordering the 70 Terabyte 2004 systems (the coefficients above the diagonal) and the 78 Robust 2003 systems (below the diagonal) according to the average metric score across topics, using the R03 and T04 topic sets, and the same weighted-by-user-T computation as was employed for Table 5. The four T-aware metrics have relatively high similarity to each other in terms of the system orderings they induce, and fit in to the middle of the spectrum, in terms of being neither deep metrics (like NDCG) nor shallow (like RR1, which is equivalent to RR). They also yield system orderings that are similar to the ordering generated by RBP0.85, for which the corresponding user model has an expected search depth of 6.7.

7. CONCLUSIONS

We have demonstrated that query variability among individuals leads to substantial changes across a range of standard relevance measures, and the effect of this source of variability is substantially more than that arising from topic or system effects. We also found that variation in expected goals of search in the number of documents (and number of queries) arises substantially from user-based factors, and is broadly correlated with increasing task complexity. Finally, we found that relevance measures that capture expectations of relevant documents and are adaptive to individual behavior are more similar to each other in terms of system orderings, and sharply dissimilar from deep metrics like AP and shallow metrics like RR.

We conclude that the aspects of variability among users regarding individual query formulation and expected goals of search can be incorporated within a batch evaluation process. The use of multiple queries per topic arising from different searchers provides a more representative characterization of the mapping from information need than just one. Systems which can perform well across such a range of queries per topic are more likely to exhibit user-generalizability. Incorporating estimates of variance due to user query-factors in a statistical power calculator would help determine the number of topics needed to reliably detect certain effect sizes.

We also suggest that the adaptive and expectation-sensitive measures we presented (especially INST) display potential in having more user-generalizability *and* more task-generalizability than existing measures, which tend to overemphasize either shallow or deep recall user behaviors. We hope to build a test collection in future to carry out more conclusive experiments on this matter.

	NDCG	AP	Q 1	INSQ	RBP 0.85	INST	RRT	ERRT	RRT 1
NDCG	–	**0.95**	**0.93**	0.83	0.81	0.79	0.79	0.78	0.68
AP	0.84	–	**0.93**	0.83	0.80	0.79	0.78	0.77	0.66
Q 1	0.80	0.81	–	0.84	0.82	0.81	0.81	0.78	0.67
INSQ	0.79	0.69	0.75	–	**0.96**	**0.96**	**0.94**	**0.92**	0.79
RBP 0.85	0.78	0.71	0.68	0.87	–	**0.95**	**0.95**	**0.92**	0.79
INST	0.77	0.67	0.74	**0.97**	0.85	–	**0.96**	**0.95**	0.82
RRT	0.78	0.67	0.76	**0.94**	0.86	**0.95**	–	**0.94**	0.81
ERRT	0.71	0.61	0.64	0.87	0.79	0.89	0.85	–	0.86
RRT 1	0.52	0.49	0.38	0.59	0.59	0.60	0.57	0.71	–

Table 6: Kendall's τ-b coefficients for the 70 Terabyte 2004 system runs ordered by the scores computed for the T04 queryset (above the lead diagonal); and for the 78 Robust 2003 system runs ordered according to the scores computed for the R03 queryset (below the diagonal). Metrics are ordered according to their τ-b coefficients relative to NDCG in the T04 comparison. Bold is for scores ≥ 0.9.

Acknowledgment This work was supported by the Australian Research Council's *Discovery Projects* Scheme (projects DP110101934 and DP140102655). We thank Alec Zwart and Xiaolu Lu.

References

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. In *SIGIR Forum*, volume 42, pages 9–15, 2008.

[2] L. W. Anderson and D. A. Krathwohl. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

[3] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proc. SIGIR*, pages 23–32, 2013.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proc. SIGIR*, pages 667–674, 2008.

[5] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. Effect of multiple query representations on information retrieval system performance. In *Proc. SIGIR*, pages 339–346, 1993.

[6] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Proc. Man.*, 31(3):431–448, 1995.

[7] D. J. Bell and I. Ruthven. Searchers' assessments of task complexity for web searching. In *Proc. ECIR*, pages 57–71, 2004.

[8] C. Buckley and J. Walz. The TREC-8 query track. In *Proc. TREC*, 1999. NIST Special Publication 500-246.

[9] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Inf. Proc. Man.*, 31(2):191–213, 1995.

[10] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.

[11] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Amer. Doc.*, 19(1):30–41, 1968.

[12] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psych. Bull.*, 76(5):378, 1971.

[13] K. Fujikawa, H. Joho, and S. Nakayama. Constraint can affect human perception, behaviour, and performance of search. In *Proc. Int. Conf. Asia-Pacific Digital Libraries*, pages 39–48. 2012.

[14] J. Gwizdka and I. Spence. What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *Proc. Amer. Soc. Inf. Sc. Tech.*, 43(1):1–22, 2006.

[15] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.

[16] G. Kazai, N. Craswell, E. Yilmaz, and S. Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proc. CIKM*, pages 105–114, 2012.

[17] D. R. Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218, 2002.

[18] G. Kumaran and J. Allan. Adapting information retrieval systems to user queries. *Inf. Proc. Man.*, 44(6):1838–1862, 2008.

[19] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. SIGIR*, pages 472–479, 2005.

[20] F. Modave, N. K. Shokar, E. Peñaranda, and N. Nguyen. Analysis of the accuracy of weight loss information search engine results on the internet. *Amer. J. Public Health*, 104(10):1971–1978, 2014.

[21] A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. AIRS*, pages 1–12, 2013.

[22] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2:1–2:27, 2008.

[23] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[24] S. E. Robertson and E. Kanoulas. On per-topic variance in IR evaluation. In *Proc. SIGIR*, pages 891–900, 2012.

[25] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *J. Inf. Ret.*, 11(5): 447–470, 2008.

[26] T. Saracevic. Relevance reconsidered. In *Proc. Conf. Conceptions of Library and Inf. Sc.*, pages 201–218, 1996.

[27] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. Lambdamerge: merging the results of query reformulations. In *Proc. WSDM*, pages 795–804, 2011.

[28] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[29] K. Spärck Jones and R. Bates. Report on the Design Study for the "Ideal" Information Retrieval Test Collection. *British Library Research and Development Report*, 5428, 1977.

[30] E. G. Toms, H. O'Brien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. Macnutt. Task effects on interactive search: The query factor. In *Focused Access to XML Documents*, pages 359–372. Springer, 2008.

[31] P. Vakkari. Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Inf. Proc. Man.*, 35(6):819 – 837, 1999.

[32] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Proc. Man.*, 36(5):697–716, 2000.

[33] W. Webber, A. Moffat, J. Zobel, and T. Sakai. Precision-at-ten considered redundant. In *Proc. SIGIR*, pages 695–696, 2008.

[34] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proc. CIKM*, pages 297–306, 2006.

[35] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proc. IIiX*, pages 254–257, 2012.

[36] W.-C. Wu, D. Kelly, and A. Sud. Using information scent and need for cognition to understand online search behavior. In *Proc. SIGIR*, pages 557–566, 2014.

[37] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. CIKM*, pages 102–111, 2006.