Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness

ALISTAIR MOFFAT, The University of Melbourne
and
PETER BAILEY, Microsoft
and
FALK SCHOLER, RMIT University
and
PAUL THOMAS, Microsoft

Information retrieval systems aim to help users satisfy information needs. We argue that the goal of the person using the system, and the pattern of behavior that they exhibit as they proceed to attain that goal, should be incorporated into the methods and techniques used to evaluate the effectiveness of IR systems, so that the resulting effectiveness scores have a useful interpretation that corresponds to the users' search experience. In particular, we investigate the role of search task complexity, and show that it has a direct bearing on the number of relevant answer documents sought by users in response to an information need, suggesting that useful effectiveness metrics must be *goal sensitive*. We further suggest that user behavior while scanning results listings is affected by the rate at which their goal is being realized, and hence that appropriate effectiveness metrics must be *adaptive* to the presence (or not) of relevant documents in the ranking. In response to these two observations, we present a new effectiveness metric, INST, that has both of the desired properties: INST employs a parameter T, a direct measure of the user's search goal that adjusts the top-weightedness of the evaluation score; moreover, as progress towards the target T is made, the modeled user behavior is adapted, to reflect the remaining expectations. INST is experimentally compared to previous effectiveness metrics, including Average Precision (AP), Normalized Discounted Cumulative Gain (NDCG), and Rank-Biased Precision (RBP), demonstrating our claims as to INST's usefulness. Like RBP, INST is a weighted-precision metric, meaning that each score can be accompanied by a *residual* that quantifies the extent of the score uncertainty caused by unjudged documents. As part of our experimentation, we use crowd-sourced data and score residuals to demonstrate that a wide range of queries arise for even quite specific information needs, and that these variant queries introduce significant levels of residual uncertainty into typical experimental evaluations. These causes of variability have wide-reaching implications for experiment design, and for the construction of test collections.

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

Additional Key Words and Phrases: User behavior, test collections, effectiveness metric, relevance measures, query, search

ACM Reference Format:
Authors. 2016. Title. *ACM Trans. Inf. Syst.* V, N, Article A (January YYYY), 37 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

This paper includes work previously published by the authors in a sequence of conference papers. Sections 3, 5, and 6, are based on work presented at the 2015 SIGIR Conference on Research and Development in Information Retrieval [Bailey et al. 2015], as is part of Section 4; the majority of Section 7 was presented as a short paper at the 2015 CIKM Conference on Information and Knowledge Management Moffat et al. [2015]; and part of Section 4 was presented as a short paper at the 2015 Australasian Document Computing Symposium [Moffat et al. 2015].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© YYYY ACM 1046-8188/YYYY/01-ARTA $15.00
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In the Cranfield and TREC evaluation frameworks, information retrieval test collections – consisting of a corpus, topics, and relevance judgments – are collectively regarded as representing a sample of some population of a real-world information retrieval task being carried out by a synthetic user. Measures of system quality, in this approach, are determined from the relevance judgments which specify, for each topic and document, how good (according to the judging guidelines) that document is for that topic, and are represented as numeric values via the use of one or more *effectiveness metrics*. This approach is sometimes referred to as *batch evaluation*. In principle, the relevance judgments cover all possible topic–document pairs, but in practice it quickly becomes infeasible to judge all documents in response to even a single query per topic, even for relatively small collections. To manage this cost, a technique known as *pooling* is used when relevance judgments are being created [Spärck Jones and van Rijsbergen 1975]. In this approach, each of the systems being measured (for example, in a shared evaluation task at an event such as TREC, NTCIR, CLEF, or FIRE) contributes a ranking of d (or more) documents, in decreasing order of predicted relevance. The union of the top-d sets is then identified, and those documents are the ones that are judged for relevance. This ensures that each system has at least its top-d documents judged (and possibly more). Extensive experimentation has shown that this approach tends to lead to an unbiased set of judgments, resulting in the same system ordering as would full evaluation [Harman 2005; Zobel 1998].

For the sake of reusability and reliability, pooling should cover possible sources of variation – for example, pools need to cover a large range of present (and anticipated) systems since different retrieval systems act differently and will retrieve different documents. The validity of scoring systems that didn't contribute to the original pool has been explored [Büttcher et al. 2007; Zobel 1998]. For smaller collections, it was found that non-contributing systems are fairly evaluated when test collections are re-used; however, for large collections, where pools are small relative to the full collection, a bias may be introduced towards those documents that directly match query keywords [Buckley et al. 2007].

Since the relevance judgments are made from a pool of documents contributed by many systems, rather than individual rankings, the method relies on assessing the documents in such a way that the judgment labels may be re-used independently. That is, judging instructions and interfaces purposely disregard any effects that the original document rank ordering of a system or document set examination strategy would induce in, or be undergone by, a motivated real user carrying out a situational task – and their cognitive changes as they do so.

One way to achieve this independence outcome is to assess the degree of an "aboutness" relationship of some (part of a) document to the topic, as per Saracevic [1996]. This approach is often characterized as "topical relevance". Another approach is that employed by TREC ad hoc assessors, where they [Voorhees 2002b]:

> *assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.*

Such "relevance" judgments are a necessary, but not sufficient, condition for a test collection to be reusable. Batch evaluation is initially only supported for systems participating in the development of the pool of judged documents. If the pool which has been constructed is sufficiently deep and diverse, it may also support batch evaluation of systems that did not contribute to the pool of judged documents.

In many modern test collections, only a single query is available per topic, and that query is provided to the system to process in order to find documents that have algorithmic relevance (that is, the system ranks them highly). However, there is no requirement that there is a single query per topic. In early TREC ad hoc test collections, manual runs were highly encouraged, since these resulted in variable queries which helped diversify document pools. Some more recent test collections, such as those associated with the TREC Million Query Track, have also sought to increase the number of topics being included, thereby increasing the number of queries overall.

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. To the extent that different topics arise from different authors, there is user-variability embedded in the test collection, but typically it is inseparable from the topic effect. It has not been standard practice for TREC test collections to provide the author identifiers of corresponding topics, such that differences among authors contributing to a test collection could be analyzed.

The effectiveness metric that is used in conjunction with the relevance judgments encodes either explicitly or implicitly an abstracted model of user behavior, and rewards systems which deliver relevant material more precisely or comprehensively according to the embodied model. Statistical assessments of comparative effectiveness as determined by the relevance measure, within the relatively small topic set and implied tasks modeled by the test collection, are then 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 [Webber et al. 2008; Sakai 2016a; 2016b]. Because the key outcome of a batch evaluation experiment is a numeric measurement of system performance averaged over the per-topic performance using an effectiveness metric which represents the aggregate relevance each answer list, batch evaluations provide a basis for repeatable longitudinal system evaluations, and enable incremental system improvements to be identified.

An important aspect of test collection use that has not been fully investigated is the degree to which they have *external validity*. Broadly speaking, external validity characterizes the extent to which an experiment – which, in the case of a test collection, measures relative system effectiveness according to some chosen effectiveness metric – generalizes to other real world circumstances. Such circumstances might encompass different users (for example, stay-at-home parents versus retired intelligence analysts) or different document sets (for example, subsets of the Web versus collections of news articles) or for different interaction tasks (for example, factoid question answering versus ad hoc topical information discovery). Potential properties of a test collection that relate to the degree to which they have external validity include: *corpus-generalizability*, whether outcomes from this test collection are consistent across other test collections modeling the same task, as investigated by Robertson and Kanoulas [2012]; *task-generalizability*, whether outcomes from this test collection are consistent across other test collections with different tasks (that is, generalizability across situations); and *user-generalizability*, whether outcomes are consistent in the presence of different user behaviors for the same task and topics (that is, generalizability across people).

More broadly, many other sources of variation are often abstracted out of the process also. For example, the usability of an interface is not considered by many widely-reported metrics, and a retrieval system is deemed to have completed its role when a ranked list of documents is produced. The interactions between a user and the list of document summaries are typically also ignored, despite their demonstrated impact on evaluation results [Turpin et al. 2009]. Indeed, the initial interaction of a user with a search system on submitting a query is with the rich and complex search results page containing many more elements than just document summaries, leading to the notion of, and evaluation method proposed for, *whole page relevance* as discussed by Bailey et al. [2010]. One notable exception to this preceding generalization is the time-biased gain measure proposed by Smucker and Clarke [2012b], which factors in overall usability through incorporating the entire time to access satisfying documents, thereby including the entire interaction process. Many sources of user variation exist (for example, age, reading level, reading speed, topical expertise, cultural background) and have been studied in small scale experiments. Systematic representation of such sources of variation within a reusable test collection would nearly always make the experimental data prohibitively expensive to obtain, and thus has been avoided.

Overall, due to the need for relevance judgments based on presentation order-independence and the complexity of modern search interfaces, the primary variables that are retained in much batch evaluation experimentation are systems and topics, with the assumption that if a sufficiently representative set of topics (and corresponding queries) is employed, system differences will be reliably identified.

In this work, we examine a limited element of the *user-generalizability* property of test collections. In particular, we seek to understand how introducing two sources of variability between users, namely individual query formulation and expectations of the quantities of relevant information needing to be found, might affect how reusable test collections are constructed, and how batch evaluations are carried out. We use the lens of *task complexity* (discussed below) to help assess these issues across a range of information seeking scenarios, and pose a series of inter-related research questions:

RQ1. Can anticipated effort and user behavior during information-seeking be incorporated in to an effectiveness metric suited for batch evaluation? If so, do changes in relative system performance emerge? (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 the existence of individual variation in initial query formulation for a single information need alter the evaluation of system performance? (Section 6.)

RQ4. To what extent are query variants adequately covered by existing pooled relevance judgments? (Section 7.)

The overarching issue we consider throughout is: to what extent do measures of system effectiveness depend on (lack of) variation in user behavior, and thus *do test collections have sufficient user-generalizability*? We are not the first to consider this issue. In an early report on the design for an ideal test collection, Spärck Jones and Bates [1977] 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 [1999] similarly concluded 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 even short queries were rarely duplicated (16%); and*
> — *systems were only somewhat variable.*

In the context of these early observations, our work in this paper can be seen as being a re-evaluation of the implications and consequences of user variation on comparative system comparison protocols.

Other forms of system experimentation are possible, including "in the laboratory" user studies based on direct observation of human subjects, and "in the wild" studies based on the inferred behavior of system users as gleaned from interaction logs. We do not consider such options in this work, but note that they do not usually generate data resources that can then be used for longitudinal and repeatable system evaluations. That is, the project described here encompasses exploration of two (among many possible) aspects of user variability in collection-based evaluations, with the goal of better connecting user experiences to batch evaluation outcomes.

2. RELATED WORK

We now present the context of previous work in which we situate our project. We consider evaluation metrics; query variability; task complexity; factors that influence search behavior; and user goals and persistence.

Evaluation Metrics. Batch evaluations rely on objective scoring of search response listings. Long-standing mechanisms include Reciprocal Rank (RR); Precision at depth k (Prec@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) [Järvelin and Kekäläinen 2002]; Rank-Biased Precision (RBP) [Moffat and Zobel 2008]; Expected Reciprocal Rank (ERR) [Chapelle et al. 2009]; and the Q-Measure (Q) [Sakai and Kando 2008]. 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 may 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 (discriminative power [Sakai 2006]), and to be just as predictive of the behavior of shallow metrics as are the shallow metrics themselves [Webber et al. 2008].

Another categorization of effectiveness metrics is based on the interpretation of the numeric metric value. *Recall-based* metrics such as AP, NDCG and Q yield scores that assess how close to the "best possible" that particular ranking is, and incorporate normalization factors derived from the total volume of relevance available in the collection for that query. *Utility-based*, or *weighted precision* metrics, yield scores that are absolute, and reflect the rate at which the user gains utility as they inspect the ranking according to the corresponding user model, regardless of how much total utility is available to be transferred.

Two important metric characteristics that we focus on in this paper are *goal sensitivity*, and *adaptivity*. Metrics that are goal-sensitive are parameterized in some manner that reflects the objective of the user in carrying out that search. In particular, suppose that a user is searching in order to build a set of T useful resources in connection with some information need (a more precise definition of T is provided in Section 3). As a simple example, Prec@k is clearly a top-weighted metric, since adjusting k corresponds to truncating the user's inspection of the ranked list at different points; and hence if (as a hypothetical) $k = T$ or $k = 2T$ is used as the evaluation depth, the truncation depth and the search goal are connected, and a (somewhat) goal-sensitive evaluation can be obtained. More usually, however, in a precision-based evaluation k is fixed across a test environment, and not varied on a per-query basis. In the same way, NDCG@k can be regarded as allowing k and T to be connected, provided that k is varied on a per-query basis; and the persistence parameter p that is used in RBP can be similarly adjusted so that on average, the user reaches an expected depth of T in the ranking before abandoning their search and taking a different action, such as reformulating their query, or making use of a different search service [Baskaya et al. 2013; Thomas et al. 2014]. Indeed, any gain-based metric can be altered so that the gain accumulated is a function not just of a particular document, but also of the total relevance of the documents that have gone before. That is, either or both of the gain and the discount might be adjusted as relevance is accumulated, so as to achieve the desired effect. In a related line of work, Carterette et al. [2012] demonstrate that any metric with a "persistence" parameter – RBP and ERR in their examples – can be considered as a distribution, rather than a point estimate, with a density that depends on the distribution of parameters across users or topics. These distributions of parameters can be learned in a single pass from log files. Smucker and Clarke [2012a] demonstrated a similar technique for the parameters of time-biased gain.

An adaptive metric is one that recognizes that the user's behavior will alter as they make progress towards their search goal. Reciprocal rank is a notable example of adaptivity. In RR, the user is assumed to continue inspecting documents in the ranking until a relevant one is found, and then end their search at that point. Expected reciprocal rank (ERR) embodies a similar user model, but with a probabilistic (rather than binary) interpretation of relevance, and expected browser utility (EBU) uses a probabilistic interpretation of document inspection and hence gain [Yilmaz et al. 2010]. On the other hand, none of Prec@k, NDCG@k, RBP, or AP are adaptive; and nor are RR, ERR, or EBU goal-sensitive.

Utility-based metrics, and the models that are their duals in terms of user behavior, are discussed in detail in Section 4, in connection with our new goal-sensitive and adaptive INST metric, a key result of this project. Moffat [2013] provides further commentary on metrics, listing seven ways in which they can be categorized in terms of their numeric properties.

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. Interactive query (re-)formulation systems are also increasingly common and have been demonstrated to assist in improving retrieval performance by (among others) Kumaran and Allan [2008]. 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 [Buckley and Walz 1999] 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 [1999] 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.

In another investigation with TREC data, Alemayehu [2003] looked at the effect of query expansion. The resulting variation in queries had a moderate effect on final effectiveness, typically a bigger effect than topic but a smaller effect than system. Similarly, results from Ferro and Silvello [2016] demonstrate that the effect of stopword choice typically dominates that of system. Other investigations of query variability in the TREC setting were shown to improve query performance through data fusion [Belkin et al. 1993; Belkin et al. 1995].

Modave et al. [2014] 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. Stanton et al. [2014] and later Palotti et al. [2015] used images of related to medical symptoms to elicit a variety of queries; however, these queries were pooled and no data is available on individual query effectiveness.

Relevance Judgment Variability. In a study examining different types and sets of judges as the source of user variability, Voorhees [2000] 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. Voorhees [2000] also observed that inter-system comparisons required more substantial differences in measure scores than for intra-system comparisons. More recently, Bailey et al. [2008] 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. In separate works, Kinney et al. [2008] and Kazai et al. [2012] confirmed that such systematic bias between different kinds of judge may exist.

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 [Vakkari 1999].

Byström and Järvelin [1995] 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 [2004], 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, Kelly et al. [2015] (see also Wu et al. [2012]) proposed a hierarchy based on the Cognitive Process Dimension of Krathwohl's Taxonomy of Learning Objectives [Krathwohl 2002], which is itself a refinement of Bloom's Taxonomy of educational objectives. Through a user study, Kelly 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 in more detail in Section 3.

Factors That Influence Searcher Behavior. Wu et al. [2014] 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. Stopping behavior was considered in more detail by Maxwell et al. [2015], who examined logs from a lab study and concluded that a simple notion of "disgust", operationalized as the number of non-relevant items seen in a list, provided the best model. Click-through behavior, being easy to observe at scale, has been the subject of much study and so-called "click models" have explained clicks on the basis of relevance, attractiveness, and simple page layout (see Chuklin et al. [2015] for a comprehensive recent survey).

The relationship between constraints and searcher behavior was studied by Fujikawa et al. [2012], 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. [2013] studied the effect of query cost on the behavior of searchers, examining the influence of alternative 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. [2008], with users showing different rates of query reformulation and page views.

In a focused study, White and Kelly [2006] 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 [2006] 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 Bell and Ruthven [2004]) 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 including experience, verbal ability, and other cognitive abilities) 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.

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, it should be sensitive to that variation. Moffat and Zobel [2008] argued that each weighted-precision metric could be interpreted as corresponding to a *user model*, a description of the behavior of the presumed user; and that the same relationship was also present in the reverse direction, so that any given user model could be regarded as a definition of a dual metric. Moffat and Zobel parameterized their suggested RBP metric with a persistence parameter p, and proposed a user model in which the first document in the result ranking is always inspected by the user; thereafter, the user is assumed to proceed from the i th document to the $i+1$ th with fixed conditional probability p. In this model – and in other weighted precision models – the metric score is the expected rate at which the user receives utility, or gain, by inspecting the ranking. Later work by Carterette et al. [2012] extended this by using distributions, rather than point estimates, of parameters such as RBP's p. Learning these distributions from logged interactions naturally results in a distribution of the final measure, from which Carterette et al. demonstrate more sophisticated understanding of performance across varied users and queries.

Rather than quantifying persistence in terms of documents, Smucker and Clarke [2012b] used time as the primary persistence factor in the model, and make use of data from a user study to calibrate their gain calculations, demonstrating that short documents take less time to read than do long ones, and that repeat documents are identified even more quickly. Jiang and Allan [2016] have also argued that user effort should be factored into retrieval evaluations, and suggest that effort is affected by the amount of gain resulting from each document, and not just by its length.

A source of disagreement between judges and users was identified by Yilmaz et al. [2014]. They found that judges are motivated by assigning a relevance label independent of the effort required to make the determination. In contrast, users will abandon examination of an individual document if the effort is considered excessive.

It has often been shown in both detailed user studies and through log analysis investigations, that users differ in simply processing both the initial search results page, and subsequent browsing behavior and associated document examination. For example, Dumais et al. [2010] found that there were at least three distinct clusters of eye gaze attention patterns within a search results page among their user study participants, encompassing people who were focused on finding just the task-completing summaries within the search results, those who focused on task-completing summaries within both the advertisements and search results, and those who exhaustively examined every summary on the page before returning to the most promising of the results. White and Drucker [2007] describe two classes of distinct and extreme user behavior based out of a long term, large scale search and browse log investigation of more than two thousand individuals. The two extreme behavior classes represented 20% of the subjects, and split into Navigators (who predominantly carried out short and rapid access to navigation targets, with low variance in their browsing behaviors) and Explorers (who displayed much greater variance in their querying and browsing behaviors).

Moffat et al. [2013] note that users may have differing expectations as to what constitutes a “successful” search, even for the same information need or even the same query, and introduce the notion of an “anticipated goal of search”, the parameter T mentioned earlier. They then use T as a parameter that shapes the probabilities assigned to documents in a weighted-precision metric, arguing that this represents a useful refinement of the simpler model for conditional continuation probability that was employed in RBP. A laboratory-based user study provided evidence to support that hypothesis. In Section 4 we build on that work, examining how user variation in queries and expected goals can be combined, and describing a user model in which the evolution of conditional continuation probabilities is adaptive and dependent on the degree to which relevance has already been accrued.

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 is structured so as 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) [Voorhees 2002a];
- **R03** Robust Track 2003, 60 topics (selected from 303–610), 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 [Voorhees 2003];
- **T04** Terabyte Track 2004, 50 topics (701–750) [Clarke et al. 2004].

This mixture of topics was selected because of their diverse origins, and because of their prior use in effectiveness experimentation, including the availability of relevance judgments.

For each topic, a *backstory* was created; a short information need statement 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. The annotators 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 1. 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, the backstories were written to speak directly to the reader, and to include hypothetical family members or friends. Figure 2 shows the original TREC presentation of one of the topics shown in Figure 1. We note that backstories owe a conceptual debt to the *simulated work task situation* developed in Borlund [2003]. Borlund’s characterization is more explicit in laying out the required attributes of the description, to include the source of the information need, the environment of the situation, the problem which has to be solved, and the objective of the search. We strongly concur with her perspectives on the value of such

Q02.1828, “What was Thailand’s original name?” Remember

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, “postmenopausal estrogen Britain” Understand

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, “Recycling successes” Analyze

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.

Fig. 1: Backstories associated with three TREC topics from different tasks in different years, together with the task type. A total of 180 information need statements were generated for a subset of topics from three TREC Tracks.

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.

Fig. 2: Original specification of Topic 734 from the TREC 2004 Terabyte Track.

“short cover stories” to both trigger individual information need interpretations in surrogate users, and to provide “the platform against which situational relevance can be judged.”

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. 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. We acknowledge that there is potential for drift between the interpretations of the backstory and the original TREC topic description or question queries that led to the TREC relevance judgments being created. As with any language translation act, this drift may be a factor in the changes we observed and discuss later in the paper. However, we believe the dominant factor in the changes arises from individual user interpretation of each backstory and their chosen query. Among other goals, our subsequent work in developing the

UQV100 collection [Bailey et al. 2016] eliminates possible discrepancies between judgments and backstories (versus topic descriptions) by creating entirely new judgments using the backstories as the situational framing for relevance assessment.

Task Complexity Levels. 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 Kelly et al. [2015], derived from a taxonomy of learning objectives presented by Anderson and Krathwohl [2001]. 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" [Kelly et al. 2015].

Based on the created backstories, each of the 180 selected TREC topics was 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.[1] The overall inter-annotator agreement among the four judges for the initial ratings was 0.664, assessed by Fleiss' κ [Fleiss 1971], 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 *Understand* 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. The three labels to the right of the example tasks in Figure 1 indicate the corresponding complexity categories that were assigned to each of the information needs.

Gathering Data. To investigate user variability in a test collection setting, an experiment was carried out using the CrowdFlower crowd-sourcing platform.[2] On signing up for the experiment, participants were first presented with an information need statement, expressed as one of the created backstories. Participants were then asked 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 via a radio-button:

- 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;

[1] To promote reproducibility, the set of 180 topic backstories and corresponding task complexity labels is available at http://dx.doi.org/10.4225/08/55D0B6A098248.

[2] http://www.crowdflower.com

Task	An information seeking activity.
Topic	A description of information content or subject matter required for some task.
Backstory	A scenario-based summary of a topic, as a statement of an information need presented to crowd workers.
Task complexity	A categorized model of cognitive information processing for some task. We use three categories: *Remember*, *Understand*, and *Analyze*.
T	Anticipated number of useful documents required to satisfy a task for some topic.
Q	Anticipated number of queries required to satisfy a task for some topic.
Q02, R03, T04	The subsets of 70, 60, and 50 topics and matching judgments, taken respectively from the TREC test collections for the 2002 Question Answering Track, the 2003 Robust Track, and the 2004 Terabyte Track.
AP, ERR, NDCG, RBP p, Q β, INSQ T, INST T	Various effectiveness metrics: Average Precision; Expected Reciprocal Rank; Normalized Discounted Cumulative Gain [Järvelin and Kekäläinen 2002]; Rank-Biased Precision (persistence parameter p) [Moffat and Zobel 2008]; Q-Measure (blending parameter β) [Sakai and Kando 2008]; Inverse Squares (user expectation parameter T) [Moffat et al. 2013]; and Adaptive Inverse Squares (user expectation parameter T) [Section 4].

Fig. 3: Definitions of significant terms and abbreviations used.

— 101+ useful pages.

Second, participants were asked: "*In total, how many different queries do you think you would need to enter to find that many useful pages?*". Answers were 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.

The first of these two variables is denoted as T, the user's anticipated relevance target; the second as Q, the anticipated number of queries (including reformulations) that would be required. Figure 3 lists these two definitions, plus a summary of other aspects of our user study and experimentation. Note that these variables represent user expectations regarding the number of documents and queries that they will need, before seeing any search results. Investigating the relationship between these estimates and actual behavior during a search is an interesting question for future work.

Third, participants were asked: "*What would your first query be?*". Answers to this question were entered in a textbox. A "submit" button then lodged the participant's answers, and took them to the next randomly selected backstory. Participants were free to complete as many topics as they liked, from one to a maximum of 180; each participant saw the backstories in a different ordering. The design of the experiment was reviewed and approved by the CSIRO Human Research Ethics Committee.

Cleaning Crowd Data. The resulting raw data set had 10,800 responses from 115 workers. Not all anonymous workers take the required tasks seriously, and where it was possible to identify clearly inappropriate responses, all of those workers' responses were removed from further analysis (the workers themselves were 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 filter removed 15 of the 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. GOAL-SENSITIVE EVALUATION

In this section we examine existing evaluation metrics, and develop the background for a user model which leads to an effectiveness metric that is both goal sensitive and adaptive, and is also intuitively appealing. Our subsequent analysis considers this metric, as well as existing metrics where appropriate, for assessing our research questions. We begin with definitions of weighted-precision metrics and three complementary specifications of user models in this context. Note that for simplicity these user models ignore a searcher's interaction with the synthesized search results page, containing summaries (and/or other information), which intermediate between the initial request and the ranked list of documents returned by the search engine. This choice means that the practice of obtaining relevance judgments solely for the documents returned by participating systems is sufficient.

Defining User Models. As part of their motivation for their rank-biased precision metric, Moffat and Zobel [2008] make use of the idea of a *user model* – a formal description of the actions of a universe of users scanning a results ranking – from which a probabilistic weighting over items in the ranking can be identified, and hence a weighted-precision effectiveness metric can be derived. For example, in RBP the probabilistic user – in effect, the average behavior over a large population of identical independent users – is modeled as always examining the first document in the ranking, and thereafter moving their attention from the one at depth i to the one at depth $i+1$ with fixed conditional probability p. The metric weightings on ranks are thus probabilities of inspection, regardless of the relevance or gain associated with the documents in those positions. Moffat et al. [2013] further develop these ideas, and introduce the notation $\mathrm{W_M}(i)$ to indicate the weight assigned to the item at depth i in the ranking by user model M, with $\sum_{i=1}^{\infty}\mathrm{W_M}(i)=1$, by definition. The metric value is then computed as

$$\sum_{i=1}^{\infty} r_i \cdot \mathrm{W_M}(i)$$

where r_i is the gain value associated with the document at depth i in the ranking, and with the metric score representing the expected rate at which relevance or gain is accrued by an average user, according to the universe captured in the model. Another way of viewing the computed metric score is as the expected gain achieved by the inspection of a single document according to the probability distribution $\mathrm{W_M}(i)$. In the case of RBP, the weighting is given by $\mathrm{W_{RBP}}(i)=(1-p)p^{i-1}$, with the first item in the ranking regarded as being at depth $i=1$.

Taking the sequential-inspection interpretation one step further, the weights $\mathrm{W_M}(i)$ can be converted to *conditional continuation probabilities*, denoted here as $\mathrm{C_M}(i)$, via the relationship [Moffat et al. 2013]

$$\mathrm{C_M}(i)=\frac{\mathrm{W_M}(i+1)}{\mathrm{W_M}(i)}\,.$$

Note that in this definition it is convenient to abbreviate the usual notation for representing conditional probabilities, and that $\mathrm{C_M}(i)$ is a localized value between zero and one, representing the "at the moment just after the i th document has been examined" choice between continuing to the $i+1$ th document in the ranking, or exiting.

Because of the connection between $\mathrm{W_M}(i)$ and $\mathrm{C_M}(i)$, any set of values $0 \le \mathrm{C_M}(i) \le 1$ can equally be used to define a user model. For example, the definition of RBP can be recast as $\mathrm{C_{RBP}}(i) = p$. Moffat et al. [2013] also define a third way of specifying a user model, via the set of probabilities that each document in the ranking is the last one examined by the sequential-inspection user:

$$\mathrm{L_M}(i) = \frac{\mathrm{W_M}(i) - \mathrm{W_M}(i+1)}{\mathrm{W_M}(1)} \, .$$

Different metrics can then be defined via the use of one (and hence all) of these three equivalent forms. For example,

$$\mathrm{C_{ERR}}(i) = (1 - r_i)$$

is a definition for Expected Reciprocal Rank (ERR) [Chapelle et al. 2009], taking the gain per document (mapped to the range zero to one) as a probability of halting at this point in the ranking. Similarly, Average Precision (AP) can be defined as (see Moffat et al. [2013, Equation 6], with errors in the subscripts corrected):

$$\mathrm{C_{AP}}(i) = \frac{\sum_{j=i+1}^{\infty}(r_j/j)}{\sum_{j=i}^{\infty}(r_j/j)} \, .$$

In the case of AP the model requires that the user has knowledge of all of the relevant documents they have not yet seen, including their positions in the ranking. This presumed reliance on documents as yet unseen by the user is one of the reasons that AP has been criticized as a metric.

Requirements for a User Model. Moffat et al. [2012] list the following five properties, slightly re-written to suit our needs here, and argue that each of these five is desirable if a user model is to be reflective of human responses to an answer ranking:

(1) The probabilities governing the model should be computable based on the properties of the prefix of the ranking that has been viewed (that is, documents that have been inspected until this point), without requiring properties of the whole collection to be established.
(2) The model should be top-weighted, with $\mathrm{W_M}(i) \ge \mathrm{W_M}(i+1)$, yet also retain non-zero weightings $\mathrm{W_M}(i)$ at all ranks i, and be, as far as possible, a smoothly varying function of i; hence, it should always have a conditional continuation probability $\mathrm{C_M}(i) > 0$.
(3) All other factors being equal, $\mathrm{C_M}(i)$ should be non-decreasing with i, that is, $\mathrm{C_M}(i) \le \mathrm{C_M}(i+1)$; conversely, the probability of the user exiting the ranking at that point should decrease with depth.
(4) All other factors being equal, the model should adapt to the presence of relevant documents in the answer ranking, so that the greater the relevance encountered by depth i, the smaller the continuation probability $\mathrm{C_M}(i)$. That is, as relevance is accumulated, the user becomes increasingly satisfied, and hence less likely to continue examining documents.
(5) The model should be goal-sensitive, and parameterized in accordance with the user's initial rationale for undertaking the search, so that, all other factors being equal, larger values of T give rise to larger values of $\mathrm{C_M}(i)$.

We acknowledge that these desiderata present an idealized situation, and that "all other factors being equal" is a caveat that might be difficult to satisfy, let alone study and measure. We also note that as yet there have been only limited user studies that would allow these requirements to be fully verified. For example, in a pragmatic sense there is unambiguous evidence that viewport size and pagination boundaries introduce notable effects on user behavior [Thomas et al. 2013], as do issues in connection with the number of irrelevant

Table I: Categorization of standard effectiveness metrics (and hence their corresponding user models) against five desired properties. Metrics Prec@k and SDCG@k (scaled DCG, with values in the range zero to one) can have their evaluation cutoff point adjusted by altering k, and RBP can have its top-weightedness adjusted via the parameter p. No user models have been developed for NDCG or the Q-Measure, and they are omitted from the table for that reason.

Metric	Prop. 1	Prop. 2	Prop. 3	Prop. 4	Prop. 5
Prec@k, SDCG@k	Yes	No	No	No	Somewhat
RR, ERR	Yes	No	No	Yes	No
AP	No	No	No	Yes	No
RBP with parameter p	Yes	Yes	Yes	No	Somewhat
Expected search length	Yes	No	No	Yes	Yes

documents encountered [de Vries et al. 2004]. And while inferred measurement of continuation probabilities has provided a level of empirical support for the relationships noted in the five desiderata, we accept that these have been based on a single experiment [Moffat et al. 2013]. Moreover, because the user has the ability to abandon the search ranking at any stage, and with relatively little additional cost, reformulate their query, there is ambiguity between behaviors that can be argued to apply to sessions, and behaviors that apply to individual queries within the session. Even so, we believe that these five principles represent a starting point for the development of plausible model-based effectiveness metrics, and that they give more precise guidance in regard to searcher behavior than do previous such models, including those associated with, for example, ERR and RBP.

Existing Metrics. Table I lists a number of current metrics, and assesses their fit against the five desiderata, including a *scaled discounted cumulative gain* metric, which is a weighted-precision model that corresponds to NDCG if it is assumed that there are at least k relevant documents in the collection. For example, Prec@k is not compliant with property 2, because $\mathrm{W}_{\mathrm{Prec@}k}(k+1)=0$; and RBP is not compliant with property 4, because $\mathrm{C}_{\mathrm{RBP}}(i)=p$ regardless of the degree of relevance accumulated through to depth i. As was noted earlier, the cutoff depth k used in Prec@k and SDCG@k can be adjusted to depend on T, the user's anticipated volume of relevance, perhaps by setting $k=2T$ on a query-by-query basis; and similarly, RBP can be made goal-sensitive by altering p, perhaps by setting $p=1-1/(2T)$. That is, with any of these three metrics we might suppose that the user is interested in viewing $2T$ documents from the ranking, and hence end up computing metric averages over topics in which different values of T have been used. (Recall that the units embodied in all weighted-precision metrics are "expected relevance accrued per document inspected", and taking the mean of such quantities is numerically sound, even if different expected evaluation depths have resulted.) In practice, however, k and/or p are usually set on a whole-of-evaluation basis, rather than adjusted on a per-query basis, hence the two "somewhat" entries in the table.

A Goal-Sensitive and Adaptive Metric. Moffat et al. [2012] (see also Moffat et al. [2013]) describe a goal-sensitive RBP-like metric they denote as INSQ, based on an inverse squares distribution (hence the name) that is directly parameterized in terms of T;

$$\mathrm{W}_{\mathrm{INSQ}}(i)=\frac{1}{S_{2T-1}}\cdot\frac{1}{(i+2T-1)^2}\,, \tag{1}$$

that is,

$$\mathrm{C}_{\mathrm{INSQ}}(i)=\frac{(i+2T-1)^2}{(i+2T)^2}\,, \tag{2}$$

where $S_k = (\pi^2/6) - (\sum_{i=1}^{k} 1/i^2)$ is a normalization constant that ensures that $\sum_{i=1}^{\infty} \mathrm{W}_{\mathrm{INSQ}}(i) = 1$. Compared to $\mathrm{C}_{\mathrm{RBP}}(i)$, which is constant, $\mathrm{C}_{\mathrm{INSQ}}(i)$ is an increasing function of i, and reflects that the more time the user has already invested in viewing a given ranking, the greater their conditional probability of stepping from rank i to rank $i+1$, thereby incorporating a sense of "sunk cost" recovery. The expected inspection depth associated with INSQ is approximately $2T + 0.5$ [Moffat et al. 2012]. In terms of the properties enumerated in Table I, INSQ is in the same category as RBP, with a "no" entry against Property 4 (adaptivity).

To develop an *adaptive* metric with some similar properties to INSQ, we now diverge from Moffat et al. [2013], and define

$$T_i = T - \mathrm{Rel}(i), \tag{3}$$

where $\mathrm{Rel}(i) = \sum_{j=1}^{i} r_i$ is the aggregate volume of relevance accumulated through to and including the document at rank i, and might be integral if the r_i values are all binary, or might be real-valued if the r_i gain values are fractional. With this formulation, T_i is an estimate of the volume of relevance still anticipated: when positive, it measures the as-yet-unmet expectation for relevance; and when negative, indicates the extent to which an excess of relevance has already been accumulated, compared to the initial user estimate. Prior to any documents being inspected, $T_0 = T$, the user's initial relevance goal. We then define a new metric INST (named as "INSQ with T", but also as an anagram of NIST, in acknowledgment of the more than two decades of support NIST have provided to IR evaluation) via

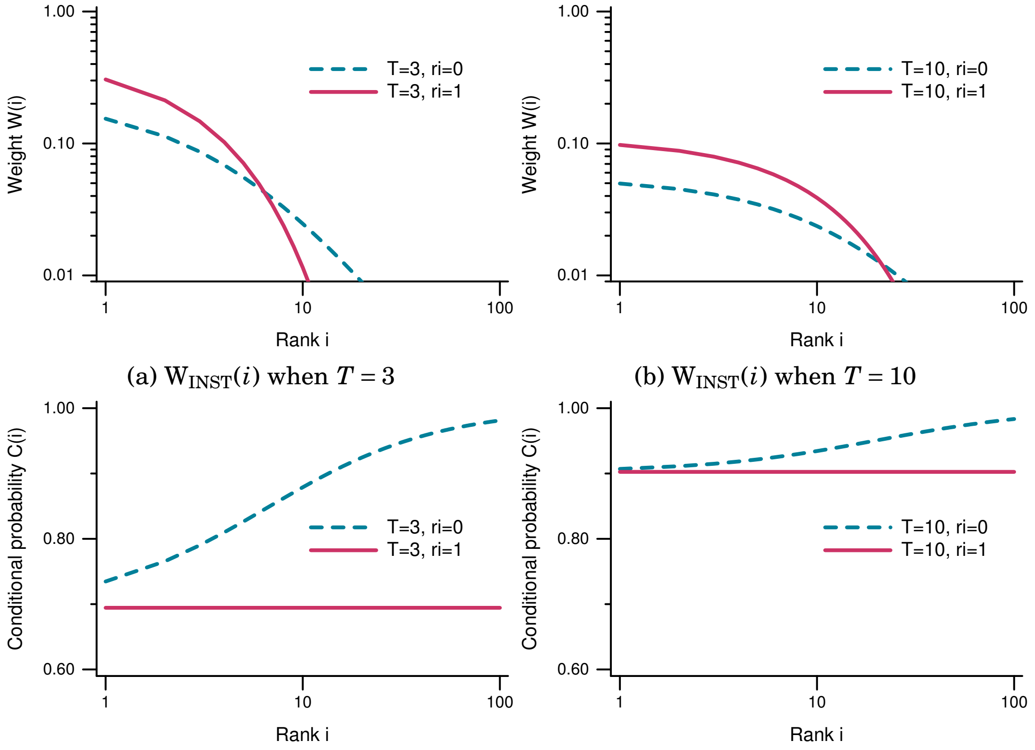
$$\mathrm{C}_{\mathrm{INST}}(i) = \frac{(i + T + T_i - 1)^2}{(i + T + T_i)^2}\,. \tag{4}$$

The motivation behind Equations 3 and 4 is simple: $\mathrm{C}_{\mathrm{INST}}(i)$ should be higher if T is higher; higher if T_i is higher; and higher if i is higher. That is, if T is high, we expect the user to proceed to a greater depth in the ranking than if T is low; if the user finds relatively few relevant documents as they scan the ranking, then T_i remains close to T and we expect them on average to proceed further than if they find an abundance of relevant documents; and underlying those two factors, we expect that the user's conditional continuation probability will increase with depth i. Note that Equation 4 is also used by Moffat et al. [2013], but with a different definition for T_i.

Figure 4 illustrates the effect of these definitions, plotting $\mathrm{W}_{\mathrm{INST}}(i)$ and $\mathrm{C}_{\mathrm{INST}}(i)$ for two different values of T. To obtain extreme cases it is assumed first that every document in the ranking is relevant, and then second, that every document is non-relevant. For a document ranking with a mix of relevant and non-relevant outcomes, the curves will then lie somewhere between these extremes. The four green dashed lines show the case when all documents are non-relevant, and are the same curves as would arise with INSQ. Conditional continuation probabilities rise with i, and approach 1 in the limit; and as a result there is a greater fraction of the weight $\mathrm{W}_{\mathrm{INST}}(i)$ shifted to deeper ranks. In terms of the corresponding model, users who do not find relevant documents are postulated as searching further down the ranking on average.

As has already been noted, some consequences of the model embedded in INST are at odds with known user behavior – for example, on a "patently bad" ranking, users are likely to abandon early and reformulate, rather than continue searching to the predicted average depth in the first result list they obtain. That is, in a "whole of session" sense, such users do indeed continue their search, but in a different list. Note too that other models for user behavior, including those associated with RBP, ERR, and AP also fail to account for this situation. Indeed, when presented with a "no relevant documents at all" ranking, both AP and ERR predict that the user will continue to inspect documents indefinitely.

On the other hand, if every document is relevant (the set of four solid red lines in Figure 4), the conditional continuation probability $\mathrm{C}_{\mathrm{INST}}(i)$ holds constant, and RBP-like behavior is anticipated. Users who find relevant documents are modeled as ending their search more



(a) $\mathrm{W}_{\mathrm{INST}}(i)$ when $T = 3$

(b) $\mathrm{W}_{\mathrm{INST}}(i)$ when $T = 10$

(c) $\mathrm{C}_{\mathrm{INST}}(i)$ when $T = 3$

(d) $\mathrm{C}_{\mathrm{INST}}(i)$ when $T = 10$

Fig. 4: The weighting function $\mathrm{W}_{\mathrm{INST}}(i)$ (upper, logarithmic vertical scale) and conditional continuation probability $\mathrm{C}_{\mathrm{INST}}(i)$ (lower) when $T = 3$ (left) and $T = 10$ (right), for two extreme rankings in which $r_i = 0$ throughout (dashed green) and $r_i = 1$ throughout (solid red).

Table II: Average ranking depth reached by users for INSQ and INST for different values of T, when no documents in the ranking are relevant (column "upper"), with both metrics having the same behavior; and when every document in the ranking is relevant (columns "lower").

T	Upper	Lower	
		INSQ	INST
1	2.58	2.58	1.33
3	6.53	6.53	3.27
10	20.51	20.51	10.26
30	60.50	60.50	30.25

quickly than users who do not, and do not become more likely to continue searching the further down the ranking they have progressed.

Expected Search Depth. Table II helps explain why we prefer INST to INSQ. It lists expected search depth – which can be computed as $1/\mathrm{W}_{\mathrm{M}}(1)$ for any weighted-precision metric M – for INSQ and INST in the two extreme situations already shown in Figure 4. 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 [Moffat et al. 2013]. On the other hand, INST is responsive to occurrences of relevant documents in the ranking, and the average depth

Table III: Kendall's τ-b coefficients for the 70 Terabyte 2004 system runs ordered by the average scores computed across the 50 topics in the T04 queryset according to different effectiveness metrics (above the lead diagonal); and for the 78 Robust 2003 system runs ordered by the average scores across the 60 topics in the R03 queryset (below the diagonal). Metrics are presented according to their τ-b coefficients relative to NDCG in the T04 comparison. Coefficients ≥ 0.9 appear in bold.

	NDCG	AP	Q 1	INSQ	RBP 0.85	INST	RR
NDCG	–	**0.95**	**0.93**	0.83	0.81	0.79	0.68
AP	0.84	–	**0.93**	0.83	0.80	0.79	0.66
Q 1	0.80	0.81	–	0.84	0.82	0.81	0.67
INSQ	0.79	0.69	0.75	–	**0.96**	**0.96**	0.79
RBP 0.85	0.78	0.71	0.68	0.87	–	**0.95**	0.79
INST	0.77	0.67	0.74	**0.97**	0.85	–	0.82
RR	0.52	0.49	0.38	0.59	0.59	0.60	–

reached is reduced if relevant documents appear. In the extreme case, if T useful documents are sought by the user, and if every document viewed is relevant, then $i + T_i$ is constant at T, and substituting in to Equation 4, we have $\mathrm{C}_{\mathrm{INST}}(i) = (2T-1)^2/(2T)^2$, which is also a constant. That is, if every document is relevant, then INST simplifies to RBP with a persistence p given by $p = (2T-1)^2/(2T)^2$. The expected ranking depth for RBP is given by $1/\mathrm{W}_{\mathrm{RBP}}(1) = 1/(1-p)$; which in the all-relevant case for INST becomes

$$\frac{1}{1-\left(\frac{2T-1}{2T}\right)^2} = \frac{1}{1-\left(1-\frac{1}{2T}\right)^2} = \frac{1}{1-\left(1-\frac{1}{T}+\frac{1}{4T^2}\right)} = \frac{T}{1-\frac{1}{4T}} \approx T\left(1+\frac{1}{4T}\right) = T+0.25\,.$$

In the all-zero case the expected depth reached in the ranking is $2T + 0.5$ [Moffat et al. 2013]. That is, if a user is seeking a total of T relevant documents, the user model corresponding to INST expects that on average they will view between $T + 0.25$ and $2T + 0.5$, with the latter occurring in a probabilistic sense only if no relevant documents appear in the ranking.

Comparing Systems. Table III lists Kendall's τ-b coefficients, computed by scoring TREC systems using a total of seven effectiveness metrics, and ordering the 70 T04 systems using average scores across 50 topics (the coefficients above the diagonal), and the 78 R03 systems using average scores across 60 topics (below the diagonal), and using the distributions of T for each topic that were generated by the crowd-workers (see Section 5). The T-aware INST metric gives similar ordering to the T-aware non-adaptive INSQ variant; both are neither deep metrics (like NDCG) nor shallow metrics (like RR). They also yield system orderings that are similar to the ordering generated by RBP 0.85, for which the corresponding user model has an expected search depth of 6.7.

Implementing INST. The adaptive nature of INST means that for any given value of i the value of $\mathrm{W}_{\mathrm{INST}}(i)$ is dependent on all values of $\mathrm{C}_{\mathrm{INST}}(i)$, and hence in turn on all values of r_i, the assessed relevance at depth i. Moffat et al. [2015] provide pseudo-code that describes the required computation. Like RBP and other weighted-precision metrics, it is possible to compute upper and lower bounds on INST scores and hence determine a *residual*, by assuming in turn that all unjudged (or unranked) documents are relevant, and non-relevant, respectively. We employ INST residuals in Sections 5 and 6. Note also that INST makes use of a per-query parameter T, and that in all of our evaluations (including, for example, in Table III) we take the distribution of values supplied by our crowd-sourced participants, and compute a weighted sum over the INST values associated with each distinct value of T.

Observations and Implications. It is difficult to show experimentally that any particular metric is "right" compared to some alternative metric, and much of the basis for believing in the usefulness of one approach compared to another comes from evidence based on observation of users as they carry out search tasks [Moffat et al. 2013], and the innate plausibility of the corresponding user model. In the case of INST, all five of the desirable requirements listed earlier in this section and tabulated in Table I are catered for, and we argue the expected search range, of $T + 0.25$ on average to $2T + 0.5$ on average, depending on the extent of relevance identified relative to the user's initial target T, is also "reasonable". We cannot, of course, provide a logical proof that INST better matches user behavior than do RBP or INSQ with suitable parameters; nor can we readily demonstrate that outcome in a beyond-doubt manner via experimental evaluations over mere dozens of users. It might be that large-scale measurement in a commercial setting – over millions of users and tens of millions of queries – is required in order to derive sufficiently delicate measurements to confirm or refute our belief that INST provides a better-fitting user model than do previous metrics such as RBP and ERR.

Nevertheless, we argue that INST is an important step forward in the evolution of effectiveness metrics, since it imbues the corresponding user model with goal-sensitivity, adaptivity, and probabilistic behavior – a combination of characteristics not previously available in a single metric. Important factors that remain outside our current framework include diversity, duplicate answers, and session-related behavior. However, a range of other work has considered these factors (see Section 2) and can, we believe, be adapted to also apply to INST.

5. VARIATION IN EXPECTATIONS

A key motivator for this work is our belief that users have differing expectations of a search system and of the information-seeking task they are engaged in, and hence that it is 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 is likely to be inappropriate to evaluate systems based only on the rank of only the first relevant result. Or, if another user expects to issue several queries in succession, then it may be appropriate to evaluate the whole session rather than measure each of the individual queries. Other, similar, scenarios are easy to imagine. Two questions in our CrowdFlower instrument were intended to understand some of these varied expectations.

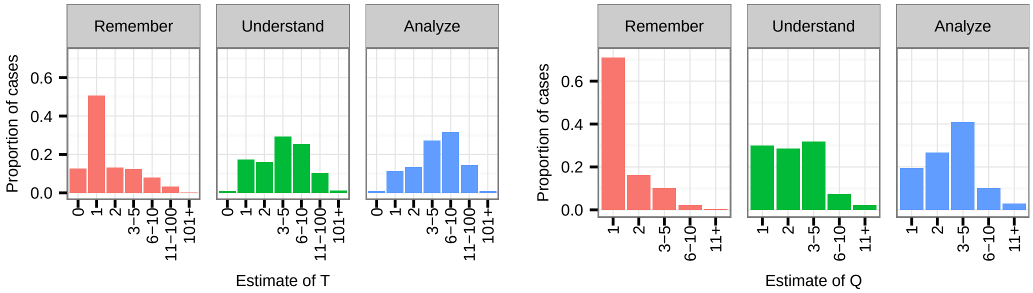
Expected Number of Useful Documents. That users commence a search with some intended goal for what they are seeking is not a new idea. Nor is the suggestion that the measurement of effectiveness should be somehow related to the extent to which that goal is met by the search. For example, it has been nearly fifty years since Cooper [1968] noted that (p.31):

> *most measures do not take into account a crucial variable: the amount of material relevant to . . . [the] query . . . which the user actually needs.*

Following Moffat et al. [2013], we denote this quantity by T. Cooper further observes (p.33):

> *The important point here is that to every search request submitted to a retrieval system there will always correspond some desired quantity of relevant material. 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, as part of the crowd-sourced elicitation described in Section 3 we asked the crowd-workers: "*How many useful web pages do you think you would need to complete the search task?*". The responses for the three task complexity categories are plotted in Figure 5(a). 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 to need to identify a greater number of useful pages.



(a) User estimates of T

(b) User estimates of Q

Fig. 5: Judges' estimates of T, the number of useful (relevant) documents they expect to need to read, vary with task complexity (left); as do their estimates of Q, the number of queries they expect to issue (right).

Estimates of T vary with task complexity, and may vary with other factors as well. Nor is T necessarily static, which is what we have assumed here. For example, discovery of a very good resource might give rise to the user re-estimating a new T and subconsciously adjusting their behavior as a result of that discovery above and beyond that accounted for by our use of T_i, including, for example, via the learning of synonyms of, or acronyms for, the original query terms. Users might also have simply made inappropriate initial estimates, and been obliged to modify their expectations as they continue the search; users will also have innately differing expectations of how much information they "need" in order to deal with the particular problem at hand. Over the universe of users, there will be a distribution of T values for a given information need.

Some of these many possible factors are captured in our CrowdFlower 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, 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.[3]

The model is summarized in Table IV, where effects are given as multipliers to odds ratios. Effects greater than one represent a higher probability of useful pages higher up the scale. For example, a multiplier of two would mean the odds of estimating T as one versus estimating T as zero are twice as high as the baseline; likewise, the odds of estimating T as two versus either one or zero 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 they expected to need to identify useful documents for only 5 out of 65 topics, and a second worker at the other extreme expecting to need 11 or more useful 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*. That is, while the identity of the worker provides the strongest predictor of the value of T, once that has been allowed for, the task type itself is the next most important factor, and in aggregate, the workers react as expected to the task complexity. Finally, there is an effect

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

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

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

due to the author of the corresponding backstory: even after controlling for task complexity and worker, some of the backstory authors provoked higher T estimates, an effect that was statistically significant, but practically negligible. We also checked for batch effects, necessary because 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 on a per-need basis. For example, if they think a task is simple or well-supported, they may be confident that a single query will suffice. On the other hand, if they anticipate that the task will be complex, they may commence the search process with the expectation that they will need to issue multiple queries. We use Q to denote the user-expected number of queries, and plot the values generated from the crowd-workers in connection with the 180 Q02, R03, and T04 backstories in Figure 5(b), and analyze their significant factors in the second data column in Table IV. As was the case already with T, the more complex the search task, the greater the value of Q, with per-judge effects dominating. Indeed, 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.

6. VARIATION IN FIRST QUERIES

As well as differing in their behavior in terms of the number of relevant answers expected, we also argue that users generate markedly different queries, even in response to the same information need, and that this is a source of uncontrolled variation in most test collection evaluations.

Normalization. The third component in each interaction pane asked "*What would your first query be?*", with workers entering text into 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 manually via spell-checker software whenever they could be unambiguously identified. For example, "theropy" was changed to "therapy" in the context of topic R03.356 (Figure 1). 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[4] were used to decide whether to alter these queries. For example, "cheapskate bay" was altered to "chesapeake bay", because that is what

[4] Google, carried out in November 2014.

Table V: 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.

	Task complexity		
	Remember	Understand	Analyze
Number of topics	70	81	29
Mean queries per topic	44.3	44.4	44.0
Mean query length (chars)	25.9	32.9	37.1
Mean query length (words)	5.6	6.0	6.5
Mean query entropy (bits)	19.9	26.0	30.5

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") that had not been caught by the quality-control mechanisms already described in Section 3 were removed. The resulting query set contained 7,969 queries, and 5,046 distinct queries.

Query Diversity. Table V lists some overall properties of those 7,969 queries, averaged over the three query classes. There is a 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 observed distribution in the usual manner [Witten et al. 1999], and finally computing the mean of 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.

As a single example, Figure 6 lists the complete set of queries generated for one of the 180 information need statements (see Figure 1). One query dominates; on the other hand, nearly half of the queries generated by the subjects occur only once. This was a typical pattern across all of the topics.

Two different retrieval computations were used to execute each query against the corresponding document collection: Indri[5] with an Okapi similarity computation [Sparck Jones et al. 2000], and Indri with a sequential dependency computation [Metzler and Croft 2005]. 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. The top 200 documents were returned and the ranking was scored; documents for which no judgment was available were deemed to be not relevant, but were noted as contributing to RBP and INST residuals.

Residuals. A risk factor in any experimentation in which judgments are re-used is the extent to which they provide coverage of the documents retrieved by the systems being compared. The *residual* associated with an RBP or INST score represents the aggregate assessment weight of all of the unjudged documents [Moffat and Zobel 2008], with zero indicating a situation in which all required judgments are available and there is no score uncertainty, and

[5] See http://www.lemurproject.org/indri/, Indri version 5.6, using Krovetz stemming, and no use of stopping.

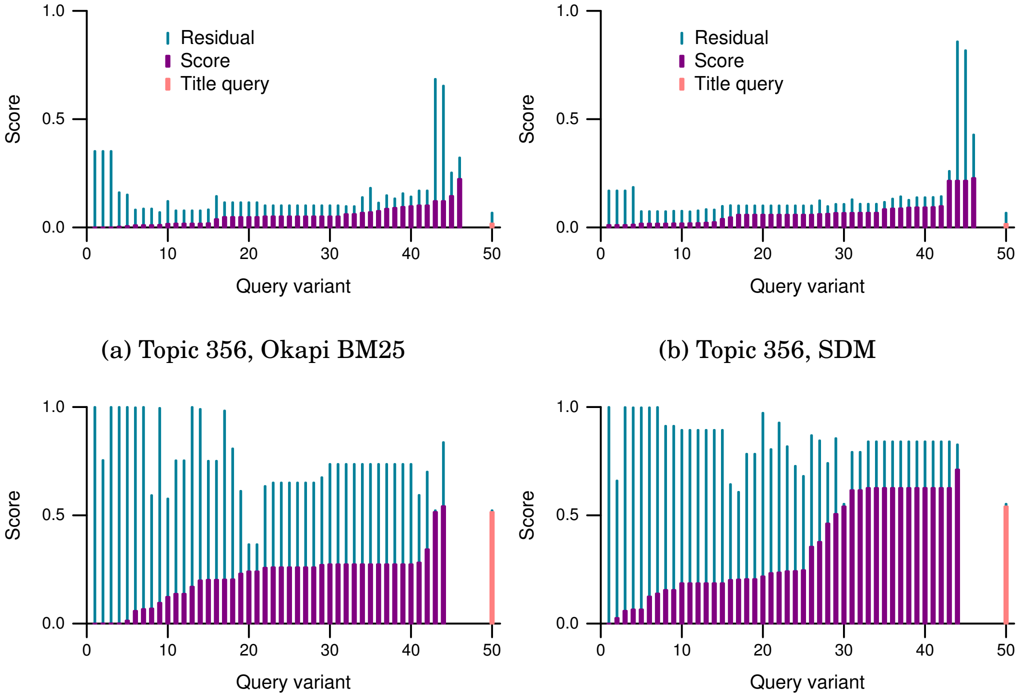
11	successful recycling projects
6	what recycling projects have been successful
2	city recycling projects
2	recycling projects
2	successful recycling programs
2	zero waste policy
1	city recycling scheme progress
1	council website
1	most successful recycling programs
1	recycling policy update
1	recycling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

Fig. 6: The 44 user-generated queries for Topic 734 (Figure 2), 25 of which are distinct. The numbers indicate multiplicity. Only one instance of the TREC title-only query "recycling successes" was generated.

values greater than zero indicating that the score would increase by that much if all currently unjudged documents were judged and found to be relevant. Figure 7 plots INST scores and residuals for the user-generated query variants for two topics, 356 and 734. For Topic 356 (Figure 7(a) and Figure 7(b)), most of the residuals were relatively low (but still comparable in magnitude to the metric scores), and only a few query variants – in particular, ones in which the crowd worker had abbreviated "hormone replacement therapy" to "hrt" – had high residuals. The omission from the corresponding backstory of the word "postmenopausal", which appears in the TREC topic description (which says "identify documents discussing the use of estrogen by postmenopausal women in Britain"), may have had an effect. As noted earlier, some level of unintentional topic drift is always possible in the process we have employed. For this topic, the TREC title-only query performs very badly, and the majority of the user variants obtain higher effectiveness scores. The best-scoring query was "hrt estrogen treatment in uk", which topped the INST base scores for both the Okapi and SDM computations.

On the other hand, for Topic 734 (Figure 7(c) and Figure 7(d)) the average INST residual for the user-generated queries (including a single occurrence of the TREC title-only query) was 0.54 for the BM25 runs and 0.50 for the SDM runs. Similar average residuals arise with RBP 0.85. On Topic 734 the title-only query was the highest-scoring query (of the 44) for AP, NDCG, and Q 1 for both Okapi and SDM models, and also the highest-scoring for RBP 0.85 for Okapi (the query "successful recycling projects place product programs" scored highest with RBP 0.85 and INST when both similarity models were used), but this topic was somewhat unusual in that regard.

Note that the TREC title-only queries in general have low residuals (for example, 0.008 and 0.012 for the BM25 and SDM runs for the title-only query for Topic 734 plotted in Figures 7(c) and 7(d) respectively), because the title-only queries were used by some of the systems that



(a) Topic 356, Okapi BM25

(b) Topic 356, SDM

(c) Topic 734, Okapi BM25

(d) Topic 734, SDM

Fig. 7: Scores and residuals when INST is used to score sets of query variants. In each graph the queries are ordered by increasing INST score. A total of 46 queries were recorded for Topic 356 (29 distinct), and a total of 44 for Topic 734 (25 distinct). In each graph the corresponding score for the TREC title-only query appears to the right of the user-generated query scores, as query "50". For Topic 356, the TREC title-only query has a very low score compared to the user query variants for both the Okapi BM25 similarity function and for the SDM computation. On the other hand, the uniformly large residuals associated with Topic 734 mean that no conclusion should be drawn either way.

contributed to the pools from which the judgments were created, and because at least some of those systems in turn will have used BM25-like and SDM-like ranking computations. (This relationship was earlier identified by Buckley et al. [2007], also in TREC pools.) The connection between judgment uncertainty and user variability, and the implications of that connection on pool construction, are examined in more detail in Section 7.

Query Effectiveness. Figure 8 provides an overview of the full sets of 60 R03 topics and 50 T04 topics, comparing the scores generated when users (and hence queries) are varied (light grey / blue boxes), and when systems are varied (dark grey / red boxes). The Q02 queries are not included because the judgment pools for those topics are relatively small and not suited to deep effectiveness metrics. To generate the system variations, the set of contributed TREC runs for the corresponding tracks were accessed from the NIST web site, and the two required query subsets extracted from them. The Indri SDM similarity scoring regime is used for the user-generated queries (blue boxes); while the red boxes make use of a wide range of similarity scoring methods, and also an (unknown) range of different queries. The average of the Indri

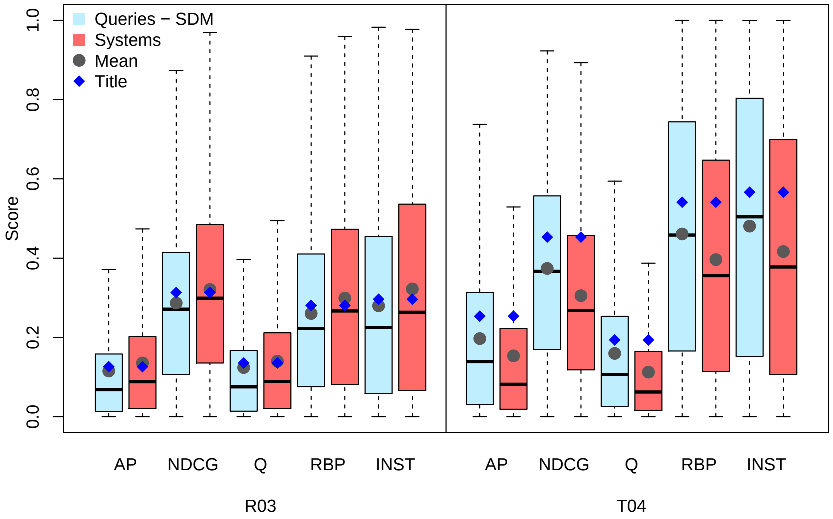


Fig. 8: Retrieval effectiveness measured by AP, NDCG, Q1, RBP0.85 and INST, for the R03 and T04 subcollections. On the left of each pair, the light grey (blue) boxes show scores obtained from all user query variants (for R03, 60 topics and 2,649 runs; for T04, 50 topic and 2,222 runs) using SDM retrieval. The darker grey (red) boxes show scores achieved by different TREC contributing system runs for the same set of topics (for R03, 60 topics, 78 systems, and 4,680 runs; for T04, 50 topics, 70 systems, and 3,500 runs). Grey points indicate the mean for each column, black bars the median, and the blue diamonds show the average effectiveness of Indri SDM runs over the same sets of 60 and 50 topics respectively, 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; and the RBP residuals for the Indri title-only SDM runs were 0.017 for R03, and 0.056 for T04.

scores for the title-only queries for the same topics is also marked on each bar. Figure 8 makes it 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 equally likely to be derived from query reformulation as it is from system improvement. Note also that for the user-generated queries (light grey / blue boxes) there is a considerable amount of metric weight still sitting in the residuals, which might be converted into increased scores if further judgments were carried out.

Figure 9 shows the extent to which the patterns depicted in Figure 7 occur overall. To construct the figure, the INST scores for the SDM mechanism for the 110 R03 and T04 topics were processed, recording the highest effectiveness score obtained by any of the user-generated topics, and plotting that best-query score as a function of the title-only SDM score for the same topic. The fraction of user-generated queries for each topic that out-scored the title-only query was also collected, and used as a further dimension in the plot. For example, 11 of the 110 topics were such that over 75% of the user-generated queries were more successful than the title-only query. There were also four topics where even the best of the user-generated queries was less effective than the title-only query, and six more where the best user-generated query equaled the score of the title-only query. On the other hand, as was demonstrated already in Figure 8, the title-only queries yield, on average, scores that are higher than the average user-generated query.

Table VI gives INST scores and residuals, and a “best” and a “worst” user-generated query for a small selection of the R03 and T04 topics, with “best” and “worst” based on the INST score

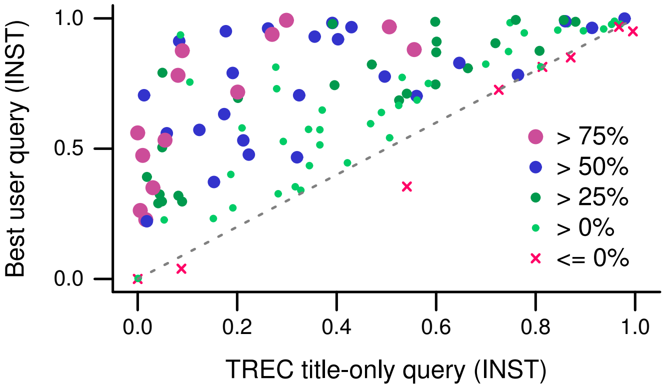


Fig. 9: Relevance scores for the best user-generated query variant for each R03 and T04 topic (110 in total) on the vertical axis, plotted as a function of the score for the corresponding TREC title-only query on the horizontal axis, measured using INST, and using SDM similarity scoring throughout. Circles represent topics where at least one of the query variants out-scored the title-only query; red crosses indicate topics where the best of the user-generated queries could only equal the title-only score, or was inferior to it. The different circle sizes and colors show the fraction of the user-generated queries for each topic that outperformed the title-only query. A non-trivial fraction of topics have sets of user-generated queries for which more than 75% of those queries score better than the TREC title-only queries, indicated by the large purple circles.

and ignoring the residual. There is one topic where the crowd-worker has pasted a sentence from the task instructions rather than generate a query, a misdirection that was not detected by our quality filtering processes. This was the only topic for which that worker had done this, and their other 179 queries were all pertinent to the corresponding information needs. What is surprising in Table VI is the quite marked score differences that arise when even seemingly small changes are made to the query. But note that there are also corresponding variations in residuals, indicating that if additional documents were to be judged, the score differences might not be as great as they seem.

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. The exact query used by each TREC system is not known to us, but by assuming it is always the same, the analysis underestimates the variability due to query phrasing and overestimates that due to system. As was noted earlier, the Q02 evaluations are quite different to the R03 and T04 ones, and are not included in this analysis.

Table VII summarizes the results. Each of query, system, and topic has a statistically significant effect ($p \ll .001$ in all cases, F-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 (for example, partial η^2 of 0.23 and 0.15 for AP), implying that slightly more variation in final score is explained by changes to system than by changes to topic. This is consistent with analysis Alemayehu [2003]. The

Table VI: Best user-generated, title-only, and worst user-generated queries (in that order) for the first five R03 and first five T04 topics, with runs generated using the Indri SDM mechanism, and scored using INST. In the case of ties for best and worst scores, one query is selected arbitrarily. For T04.705, the worst query represents a failure of our quality filtering processes (the "query" is text pasted from elsewhere on the crowd-sourcing page).

Topic	Score + residual	Query
R03.303	0.354 + 0.027	achievements of hubble telescope
	0.317 + 0.026	hubble telescope achievements
	0.042 + 0.032	hubble telescope discoveries
R03.307	0.747 + 0.083	hydroelectric power projects world
	0.599 + 0.016	new hydroelectric projects
	0.050 + 0.802	can hydroelectric energy as a green alternative
R03.310	0.467 + 0.068	brain cancer link with radio tower
	0.320 + 0.060	radio waves and brain cancer
	0.000 + 0.518	use of mobiles in cars
R03.314	0.561 + 0.043	algae seaweed kelp nutrition medicine
	0.000 + 0.143	marine vegetation
	0.000 + 0.420	marine vegetation health benefits
R03.320	0.705 + 0.035	the flag fiber optic link around the globe wiki
	0.013 + 0.044	undersea fiber optic cable
	0.001 + 0.999	who is behind the flag
T04.701	0.327 + 0.419	what factors have shaped the u s based oil industry
	0.282 + 0.015	u s oil industry history
	0.002 + 0.981	oil exploration us
T04.702	0.964 + 0.013	definition of a cultured pearl is and how they are formed
	0.913 + 0.003	pearl farming
	0.096 + 0.543	about pearls
T04.704	0.774 + 0.190	green party goals and views
	0.532 + 0.035	green party political views
	0.004 + 0.895	u s green party
T04.705	0.983 + 0.016	international aide for reducing iraqi debt
	0.749 + 0.006	iraq foreign debt reduction
	0.000 + 1.000	*the web pages that are returned by the search engine fall in to two categories*
T04.706	0.791 + 0.131	ways of keeping triglycerides cholesterol and blood pressure in normal ranges
	0.191 + 0.037	controlling type ii diabetes
	0.000 + 0.999	being healthy

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 values in the range 0.50–0.57). This is more variation than seen by Alemayehu [2003] or Ferro and Silvello [2016], which may be because our queries (from different people) vary more than did theirs (from different stopword lists, or with and without query expansion). Banks et al. [1999] observed a larger system-topic interaction effect in their comprehensive analysis of early TREC ad hoc data; it may be that we fail to observe similar effects because the query variations we analyze make use of only two systems (Indri variants) while the various TREC systems only (we assume) process a single query variation.

Observations and Implications. This section has demonstrated that query variability among individuals leads to substantial changes across a range of standard relevance measures, and that the effect of this source of variability is substantially more than that arising from topic or system effects. Particular choices of query 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 (for example, "system B is better"); less commonly, we use variation in scores to say something about topics (for example, "Topic 356 is hard"). In either

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

Metric		η^2	SS	df	F
AP	query	0.53	152.26	4894	4.32
	system	0.23	39.24	149	36.55
	topic	0.15	23.26	178	15.90
NDCG	query	0.57	270.20	4894	5.17
	system	0.30	84.37	149	52.98
	topic	0.16	38.16	178	20.06
Q1	query	0.56	142.03	4894	4.88
	system	0.20	28.54	149	32.20
	topic	0.18	24.18	178	22.84
RBP0.85	query	0.51	328.77	4894	4.00
	system	0.21	84.09	149	33.62
	topic	0.13	47.53	178	15.90
INST	query	0.50	448.30	4894	3.90
	system	0.18	93.40	149	26.69
	topic	0.13	63.64	178	15.22

case we need to be aware that the actual query wording is a significant confound, and a source of variation which in fact dominates the variation due to system and the variation due to topic.

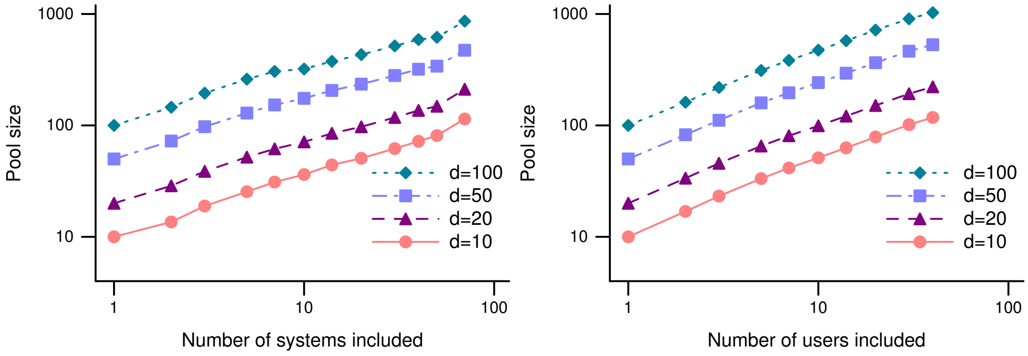
Two macro implications can be drawn from this analysis regarding test collection design and development. First, since the query-elicitation approach we have described here supplies between one and two orders of magnitude more queries for a given set of topics within a collection than do previous approaches, judgment budgets can be expected to sharply increase, even assuming some cross-query document overlap within each topic. Section 7 returns to this issue and quantifies that cost.

Second, systems could be provided with each topic's collection of queries, and 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 [Sheldon et al. 2011] could be explored.

7. VARIATION AND POOLING

Our goal in this section is to measure the extent to which the user-generated queries for any particular information need retrieve documents that are not fetched by the canonical TREC title-only query. We again omit topics from the 2002 Question Answering Track, primarily because the NIST judging process for them was oriented towards identification of facts rather than ad hoc relevance labels. To measure pool overlap, each query variant was executed using the two retrieval mechanisms already employed in Section 6: Indri's Okapi BM25 implementation, and Indri's SDM mechanism. The sets of answer rankings for each topic were then compared using two different methods, seeking in each case to quantify the extent to which the rankings overlapped.

In the first comparison approach, we compute the size of the pool of documents that would need to be judged if all variant user queries were to be evaluated fairly by a depth-d effectiveness metric, regardless of whether or not the TREC title-only query is present in the set. For example, if $d = 10$, the set of 46 queries (29 distinct) generated for R03.356 can place anywhere between 10 and 290 documents into the judgment pool, varying from uniformly perfect overlap through to uniformly disjoint orderings. In fact, with SDM evaluation, a total



(a) Pool size as systems are added

(b) Pool size as users are added

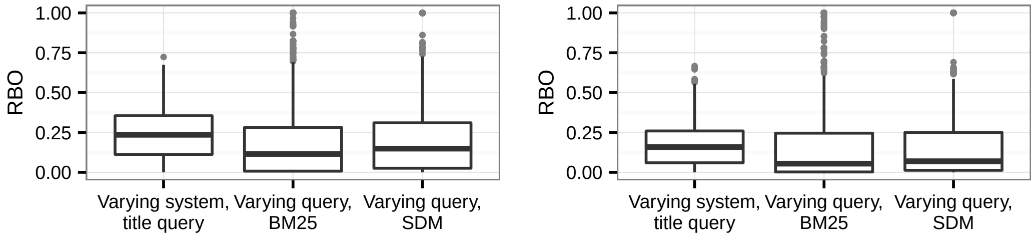
Fig. 10: Pool size (average documents per topic) as a function of systems (left) and users (right), using 60 topics from the TREC 2003 Robust Track, and a range of pooling depths d. The "systems" are TREC runs, ordered and included incrementally in the pool alphabetically; the "users" are crowd-workers, ordered and included incrementally in the pool according to their CrowdFlower ID. The set of workers who created queries differs from topic to topic; in total, 92 workers contributed an average of 28.7 queries each.

of just 45 distinct documents occur in the $d = 10$ rankings; and at $d = 20$, the corresponding number is 90. These relatively small pool sizes are encouraging; on the other hand, for T04.734 the two pool sizes (44 user-generated queries, 25 distinct) are 178 and 316 respectively.

The second comparison approach takes the TREC title-only query as a reference point, and computes the average divergence from that document ranking. To give appropriate weighting to the more highly ranked documents, and to allow for the fact that the pairs of lists are not necessarily permutations of each other (a requirement that is a necessary precondition to computing Kendall's τ, a commonly used non-top weighted correlation measure), we employ the *Rank-Biased Overlap* computation (RBO) described by Webber et al. [2010]. Rank-Biased Overlap quantifies the expected overlap observed by a probabilistic user who compares the two lists, advancing from depth d' to depth $d'+1$ with probability p. When $p = 0.95$, the value used in our experiments, the expected depth reached is $d' = 20$, but the measure is top weighted, and the greatest single contribution (with weight 0.05) arises at depth $d' = 1$, depending entirely on whether the two lists have the same first element. Hence, a high value of RBO indicates that the two lists are similar, at least to a user with the level of persistence p; conversely, low values of RBO indicate dissimilar or non-overlapping orderings.

We use the TREC title-only query as the reference point purely because it represents a simple starting position; indeed the titles for the Robust02 track topics were typically length limited to just three keywords. Any query could have been used as an anchor point for each topic; or all pairwise RBO scores could have been averaged. Our use of the TREC title-only queries is primarily in deference to more than two decades of TREC-sponsored batch-mode retrieval experimentation; but also because these are the queries that have the most complete coverage in the TREC judgments created in the past.

Pool Size. Figure 10 shows how total pool size grows as a function of the number of systems contributing to the pool for each topic (Figure 10(a)), and of the number of users contributing to the pool for each topic (Figure 10(b)), in both cases averaged across the 60 selected R03 topics. The "systems" in the left-hand graph are the 78 runs submitted by the TREC participants, ordered according to the alphabetical run name, with the same ordering used for each topic. In the right-hand graph, the horizontal axis represents queries as generated by "users" (crowd-workers) based on the information need statements; they are ordered by workers' unique



(a) RBO scores for 60 R03 topics

(b) RBO scores for 50 T04 topics

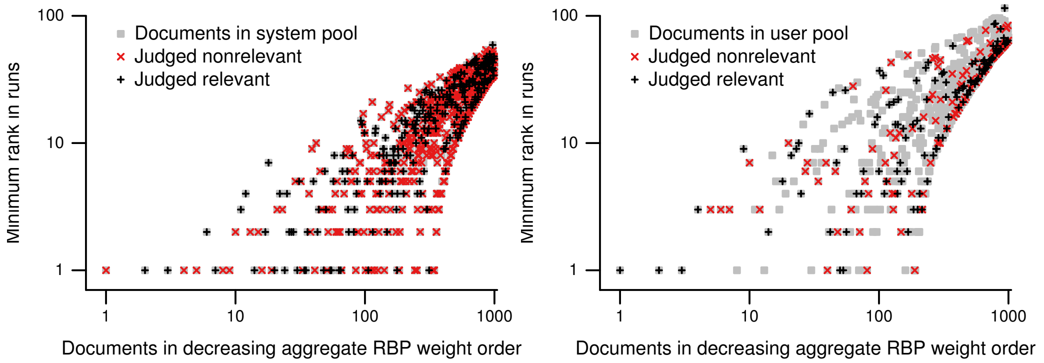
Fig. 11: RBO comparisons between ranked lists for the TREC 2003 Robust Track (left), and the TREC 2004 Terabyte Track (right).

CrowdFlower ID. Repeat queries are included in the counts, even though they do not increase the pool size. If only distinct queries were counted, the lines in Figure 10(b) would have somewhat steeper gradients. Note also that the pool of users varies across topics, because most of the workers only processed a subset of the 60 selected topics, but that this doesn't alter the conclusions that can be drawn from the graph.

The trend in Figure 10(b) is clear: for each different pool depth d, the growth in total pool size follows the same trends as it does for systems. Both options give rise (very broadly) to straight lines in the plotted log-log graphs, but with slightly different exponents: in the case of systems, the four lines are approximated by $v \approx dn^{0.5}$, where v is the volume of judgments, d is the pool depth, and n is the number of systems; in the case of users, the growth rate is higher, with $v \approx dn^{0.7}$. That is, for these users, and these systems, the pools that arise from a given number of users are larger than the pools that arise from the same number of systems.

Ranked List Similarity. Figure 11 shows the RBO scores for system-versus-system and user-versus-user comparisons for the TREC 2003 Robust Track and the TREC 2004 Terabyte Track, where the higher the RBO score, the greater the weighted overlap between the two rankings being compared. In this experiment the “systems” are TREC contributed runs that self-identified as having made use of title-only queries. In the case of the R03 collection, there were 8 runs used, across 60 topics and hence $60 \times 8 \times 7/2 = 1{,}680$ RBO scores computed. In the case of the T04 dataset, we used the eight highest scoring title-only runs as identified by the Track organizers [Clarke et al. 2004, Figure 6], across 50 topics, and hence 1,400 RBO scores were generated. To compare “users”, the sets of user-generated queries for each topic were evaluated using Indri's Okapi BM25 and SDM computations, and then those runs compared against the TREC title-only run for that topic using the same computation. For R03, this yields 2,649 RBO scores; for the T04, there are 2,222 RBO scores generated. Figure 11 plots the resulting RBO distributions. In both cases a one-sided Mann-Whitney U test indicates that the “varying user” RBO scores were significantly lower than the “varying system” scores, at $p < 0.001$, further supporting our contention that user variability is at least as high as system variability.

Strategic Pooling. Moffat et al. [2007] suggest that if a limited judgments budget is available, documents in the pool could be ordered according to how much they reduce the total amount of imprecision in the measurement system as a whole. They do this in the context of weighted-precision effectiveness metrics – Rank-Biased Precision (RBP) in particular – by summing the “votes” for each document, and if J judgments can be carried out, identifying the J documents with the largest sums of weighted votes. The selection can be done on a topic-by-topic basis, or on a global all-topics basis.



(a) document pool built from systems (b) document pool built from user queries

Fig. 12: Depth of judged and unjudged documents for Topic T04.734, with pooling by systems (left), and users (right). The horizontal axis orders documents by their summed RBP-like assessment weights when evaluated using the SDM similarity computation; the vertical axis shows the minimum rank of that document in any of the runs. The $d = 200$ pool for the 70 systems contains 4,356 documents; the $d = 200$ pool for the 44 user-generated queries (25 distinct) for this topic contains 2,570 documents. The first 1,000 documents ordered by aggregate RBP weight are plotted in each case.

Figure 12 shows how the 1,437 TREC relevance judgments for topic T04.734 would sit with this approach. As before, each of the user query variations was evaluated using the Okapi SDM computation used by Indri, and a run of 200 documents created. Each document in each run was then assigned a weight depending on its position in the run, with the document at rank r given a weight of $(1 - p)p^{r-1}$, with persistence parameter $p = 0.95$. The total weight of each document was the sum of the weight accorded that document in the 44 runs generated from the query variations, including repeats, or in the 70 system variations. To create the document ordering shown in the horizontal axes in the figure, the pool was then ordered by decreasing total weight, and plotted against minimum rank position, that is, the shallowest depth in any of the corresponding runs at which that document appeared. The correlation that shows is a consequence of the minimum RBP weight associated with each rank. A gray square is plotted for each document. The TREC relevance judgments are then overlaid – red crosses for the 1,032 documents judged as non-relevant, and black plus signs for the 405 documents judged and identified as being relevant (within the first 1,000 documents within the pool).

In Figure 12, a depth-based pooling mechanism can be thought of as generating a set of documents below a given horizontal line at rank d drawn across the graphs, with all documents below it added to the pool; and the strategic pooling mechanism of Moffat et al. [2007] as being a vertical line that sweeps from left to right, taking documents until the given judgment budget is allocated. Because the NIST-provided relevance judgments are derived from a depth-based pool built from the system runs, they provide a high level of coverage when the set of run variations is the set of systems (Figure 12(a)). But in Figure 12(b), the NIST-provided relevance judgments provide relatively poor coverage of the documents selected by the query variants, even at relatively shallow pool depths, and irrespective of the ordering applied to the documents. This mismatch is reason for the high residuals for this topic that were noted in Figure 7.

Tables VIII and IX detail the relationship between the system-generated pools for the selected topics, the user-generated pools (using the Indri Okapi similarity rankings), and the existing TREC relevance judgments. In both data sets the user-generated pools have many more unjudged documents than the system-generated pools, further evidence that user

Table VIII: Fraction of available pool covered by existing judgments, averaged over 60 TREC 2003 Robust Track topics evaluated using the BM25 retrieval mechanism. The judgments for these topics were created in 1998, 1999, 2000, and 2003, with the 2003 runs not re-judged against the earlier topics [Voorhees 2003].

Source	Depth	Size	Rele.	Irre.	Unjd.
TREC systems	$d = 20$	216.0	11.3%	74.0%	14.7%
(78 runs)	$d = 100$	883.1	5.5%	61.5%	32.9%
Users	$d = 20$	240.5	10.3%	45.8%	43.9%
(44.2 avg.)	$d = 100$	1110.9	4.5%	32.3%	63.3%
Combined	$d = 20$	394.7	8.5%	56.3%	35.2%
(122.2 avg.)	$d = 100$	1701.2	3.6%	39.2%	57.2%

Table IX: Fraction of available pool covered by existing judgments, averaged over 50 TREC 2004 Terabyte Track topics evaluated using the BM25 retrieval mechanism. The original judging for these topics covered a subset of the 70 runs submitted to the Track, and to a depth of $d = 85$ [Clarke et al. 2004].

Source	Depth	Size	Rele.	Irre.	Unjd.
TREC systems	$d = 20$	445.7	22.4%	63.0%	14.6%
(70 runs)	$d = 100$	1912.9	11.6%	49.0%	39.4%
Users	$d = 20$	236.7	25.4%	28.6%	46.0%
(44.4 avg.)	$d = 100$	974.7	14.3%	22.0%	63.7%
Combined	$d = 20$	607.7	19.3%	51.6%	29.1%
(114.4 avg.)	$d = 100$	2527.8	9.1%	38.3%	52.6%

variations give rise to quite different sets of documents being retrieved. In the rows labeled "combined", all system runs and all Okapi-based user runs are combined in a single pool. The fact that the combined pools are close in size to the sum of the two separate pools highlights the relatively disjoint nature of the respective sets of documents. Examining the combined pools, approximately one third of documents are unjudged at a pool depth of 20, and more than half are unjudged at depth 100.

In recent work we have exploited the mechanisms developed here, and used crowd-workers to generate relevance labels for more than 55,000 documents generated in response to 100 further backstories, with an average of nearly 50 distinct queries per topic [Bailey et al. 2016]; Moffat [2016] uses that collection to further examine judgment pool growth.

Observations and Implications. Our findings can be summarized simply: diversity in the pools of documents for judging arising from user variation is at least as substantial as that from system variation. We have also demonstrated that pools grow in size at a broadly comparable rate in each case, as functions of the number of systems or number of users. Moreover, even when using the same system, different user queries typically produce very different ranked lists. These results make it clear that incorporating user variation cannot be a post-hoc exercise, after a collection has been created in a conventional manner using variation solely from systems. There are substantially larger sets of unjudged documents from new "user runs" than existing "system runs", and these can occur even in the very early part of the corresponding runs. And while we cannot know whether there are relevant documents in those regions without actually carrying out judgments on them, nor is it prudent to assume that there are not – we need to accept that we simply do not know. Bailey et al. [2016] and Moffat [2016] explore this issue further.

Given finite budgets, this implies that measures that accumulate and report missing judgments via a residual score, such as RBP or INST should be used, even if solely as an adjunct to other reported metrics to allow confidence in their scores [Lu et al. 2016]; or that more cost-effective judgment acquisition methods such as crowd-sourcing approaches (for example, as discussed by Alonso et al. [2008]) should be employed.

8. CONCLUSION

In test collection-based evaluation of IR systems, variations in user behavior are often abstracted out of the measurement process. While this may increase the sensitivity of the evaluation to changes in underlying search algorithms, it can mask the impact that search system changes will have on actual users. In this work we have examined the user generalizability of test collections by considering the impact of individual query variation – the fact that different users may enter different queries for the same underlying information need – and user expectations – the amount of information that a user thinks they will need to accumulate to address their information need.

Given the existence of user variability, the first research question asked how user expectations can be incorporated into an effectiveness metric, and whether this leads to changes in relative system effectiveness. We have described a new metric, INST, that is both goal-sensitive and adaptive. We believe that INST is the first metric to combine these two desirable properties since Cooper's expected search length, and the only metric to satisfy all five desiderata of Table I.

The impact of variation of anticipated effort among users was the second research question. Our experiments demonstrated a significant level of individual variation regarding searcher expectations, as well as a direct relationship between a user's anticipated effort – both in terms of the number of documents that they expect to need to view, and the number of queries they expect to need to issue – and information task complexity. This suggests that one way to increase user generalizability of test collections may be to incorporate user expectations into an evaluation metric.

The third research question considered whether substantial individual variation in initial query formulation for a single information need affects the evaluation of system performance. Our experiments demonstrated that query variability can lead to substantial changes in effectiveness outcomes for standard evaluation metrics, and that query variability often exceeds variability arising from system or topic effects. The presence of user variability should therefore be considered as a potential confound when assessing the impact of algorithmic changes based on evaluation using test collections.

With substantial levels of user variability in evidence, an important consideration is the extent to which query variations are adequately covered by existing pooled relevance judgments, the final research question. Analysis of topics from the TREC 2003 Robust Track and 2004 Terabyte Track showed that the range of documents that are contributed to the pool due to user variation is at least as extensive as the range of documents that are added due to system variation, with the pool size growing at comparable rates for both cases. Variable user queries typically lead to very different answer lists, even when run using the same retrieval system; it is thus even more important that evaluation metrics should be reported that enable the quantification of the uncertainty that arises as a residual score.

Incorporating user variation into existing test collections is difficult, since large numbers of new documents are returned. We have recently taken steps to rectify that problem [Bailey et al. 2016], by constructing a test collection with query variation from users incorporated at design time. This new "UQV100" test collection[6] includes 100 fresh backstories, a set of 10,835 user-generated query variations, and pooled relevance judgments relative to those variations and backstories. We hope to use the UQV100 collection to help further understand the important questions we have considered in this paper.

[6] Available from dx.doi.org/10.4225/49/5726E597B8376.

ACKNOWLEDGMENTS

We thank the Associate Editor and the three anonymous referees for their detailed and helpful comments. Xiaolu Lu (RMIT University) provided invaluable assistance during this project, and Bevan Koopman has developed an `inst_eval` implementation that can be used to compute INST scores, see https://github.com/ielab/inst_eval. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (project DP140102655).

REFERENCES

N. Alemayehu. 2003. Analysis of performance variation using query expansion. *Journal of the American Society for Information Science and Technology* 54, 5 (2003), 379–391.

O. Alonso, D. E. Rose, and B. Stewart. 2008. Crowdsourcing for Relevance Evaluation. *SIGIR Forum* 42, 2 (2008), 9–15.

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

L. Azzopardi, D. Kelly, and K. Brennan. 2013. How Query Cost Affects Search Behavior. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 23–32.

P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter?. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 667–674.

P. Bailey, N. Craswell, R. W. White, L. Chen, A. Satyanarayana, and S. M. M. Tahaghoghi. 2010. Evaluating Whole-Page Relevance. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 767–768.

P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2015. User Variability and IR System Evaluation. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 625–634.

P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 725–728. Public data: http://dx.doi.org/10.4225/49/5726E597B8376.

D. Banks, P. Over, and N.-F. Zhang. 1999. Blind men and elephants: Six approaches to TREC data. *Information Retrieval* 1, 1-2 (1999), 7–34.

F. Baskaya, H. Keskustalo, and K. Järvelin. 2013. Modeling Behavioral Factors in Interactive Information Retrieval. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 2297–2302.

N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. 1993. Effect of Multiple Query Representations on Information Retrieval System Performance. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 339–346.

N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. 1995. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management* 31, 3 (1995), 431–448.

D. J. Bell and I. Ruthven. 2004. Searchers' Assessments of Task Complexity for Web Searching. In *Proc. European Conf. in Information Retrieval (ECIR)*. 57–71.

P. Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research* 8, 3 (2003).

C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. 2007. Bias and the Limits of Pooling for Large Collections. *Information Retrieval* 10, 6 (2007), 491–508.

C. Buckley and J. Walz. 1999. The TREC-8 Query Track. In *Proc. Text Retrieval Conf. (TREC)*. NIST Special Publication 500-246.

S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. 2007. Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 63–70.

K. Byström and K. Järvelin. 1995. Task Complexity Affects Information Seeking and Use. *Information Processing & Management* 31, 2 (1995), 191–213.

B. Carterette, E. Kanoulas, and E. Yilmaz. 2012. Incorporating variability in user behavior into systems based evaluation. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 135–144.

O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 621–630.

A. Chuklin, I. Markov, and M. de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool.

C. L. A. Clarke, N. Craswell, and I. Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *Proc. Text Retrieval Conf. (TREC)*.

W. S. Cooper. 1968. Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. *American Documentation* 19, 1 (1968), 30–41.

A. P. de Vries, G. Kazai, and M. Lalmas. 2004. Tolerance to Irrelevance: A User-Effort Evaluation of Retrieval Systems without Predefined Retrieval Unit. In *Proc. Recherche d'Information etses Applications (RIAO)*. 463–473.

S. T. Dumais, G. Buscher, and E. Cutrell. 2010. Individual differences in gaze patterns for web search. In *Proceedings of the third symposium on Information Interaction in Context*. ACM, 185–194.

N. Ferro and G. Silvello. 2016. A General Linear Mixed Models Approach to Study System Component Effects. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 25–34.

J. L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psych. Bull.* 76, 5 (1971), 378.

K. Fujikawa, H. Joho, and S. Nakayama. 2012. Constraint Can Affect Human Perception, Behaviour, and Performance of Search. In *Proc. International Conf. Asia-Pacific Digital Libraries (ICADL)*. 39–48.

J. Gwizdka and I. Spence. 2006. What Can Searching Behavior Tell Us About the Difficulty of Information Tasks? A Study of Web Navigation. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2006), 1–22.

D. K. Harman. 2005. The TREC Test Collections. In *TREC: Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman (Eds.). MIT Press, Chapter 2, 21–52.

K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. on Information Systems* 20, 4 (2002), 422–446.

J. Jiang and J. Allan. 2016. Adaptive Effort for Search Evaluation Metrics. In *Proc. European Conf. in Information Retrieval (ECIR)*. 187–199.

G. Kazai, N. Craswell, E. Yilmaz, and S. M. M. Tahaghoghi. 2012. An Analysis of Systematic Judging Errors in Information Retrieval. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 105–114.

D. Kelly, J. Arguello, A. Edwards, and W.-C. Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *Proc. International Conf. on Theory of Information Retrieval (ICTIR)*. 101–110.

K. A. Kinney, S. B Huffman, and J. Zhai. 2008. How Evaluator Domain Expertise Affects Search Result Relevance Judgments. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 591–598.

D. R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218.

G. Kumaran and J. Allan. 2008. Adapting Information Retrieval Systems to User Queries. *Information Processing & Management* 44, 6 (2008), 1838–1862.

X. Lu, A. Moffat, and J. S. Culpepper. 2016. The Effect of Pooling and Evaluation Depth on IR Metrics. *Information Retrieval* 19, 4 (2016), 416–445.

D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 313–322.

D. Metzler and W. B. Croft. 2005. A Markov Random Field Model for Term Dependencies. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 472–479.

F. Modave, N. K. Shokar, E. Peñaranda, and N. Nguyen. 2014. Analysis of the Accuracy of Weight Loss Information Search Engine Results on the Internet. *Amer. J. Public Health* 104, 10 (2014), 1971–1978.

A. Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Proc. Asia Information Retrieval Societies Conf. (AIRS)*. 1–12.

A. Moffat. 2016. Judgment Pool Effects Caused by Query Variations. In *Proc. Australasian Document Computing Symp. (ADCS)*. 65–68.

A. Moffat, P. Bailey, F. Scholer, and P. Thomas. 2015. INST: An Adaptive Metric for Information Retrieval Evaluation. In *Proc. Australasian Document Computing Symp. (ADCS)*. 5:1–5:4.

A. Moffat, F. Scholer, and P. Thomas. 2012. Models and Metrics: IR Evaluation as a User Process. In *Proc. Australasian Document Computing Symp. (ADCS)*. 47–54.

A. Moffat, F. Scholer, P. Thomas, and P. Bailey. 2015. Pooled Evaluation Over Query Variations: Users are as Diverse as Systems. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 1759–1762.

A. Moffat, P. Thomas, and F. Scholer. 2013. Users Versus Models: What Observation Tells Us About Effectiveness Metrics. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 659–668.

A. Moffat, W. Webber, and J. Zobel. 2007. Strategic System Comparisons via Targeted Relevance Judgments. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 375–382.

A. Moffat and J. Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. on Information Systems* 27, 1 (2008), 2.1–2.27.

J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. F. Jones, M. Lupu, and P. Pecina. 2015. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information About Medical Symptoms. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF)*.

S. E. Robertson and E. Kanoulas. 2012. On Per-Topic Variance in IR Evaluation. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 891–900.

T. Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 525–532.

T. Sakai. 2016a. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 5–14.

T. Sakai. 2016b. Topic set size design. *Information Retrieval* 19, 3 (2016), 256–283.

T. Sakai and N. Kando. 2008. On Information Retrieval Metrics Designed for Evaluation With Incomplete Relevance Assessments. *Information Retrieval* 11, 5 (2008), 447–470.

T. Saracevic. 1996. Relevance Reconsidered. In *Proc. Conf. Conceptions of Library and Information Science (COLIS)*. 201–218.

D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. 2011. LambdaMerge: Merging the Results of Query Reformulations. In *Proc. Conf. on Web Search and Data Mining (WSDM)*. 795–804.

M. D. Smucker and C. L. A. Clarke. 2012a. Modeling user variance in time-biased gain. In *Proc. Symp. Human-Computer IR*. 1–10.

M. D. Smucker and C. L. A. Clarke. 2012b. Time-Based Calibration of Effectiveness Measures. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 95–104.

K. Spärck Jones and R. G. Bates. 1977. *Report on the Design Study for the "Ideal" Information Retrieval Test Collection*. Technical Report 5428. Computer Laboratory, University of Cambridge. British Library Research and Development Report.

K. Spärck Jones and C. J. van Rijsbergen. 1975. *Report on the Need For and the Provision Of An "Ideal" Information Retrieval Test Collection*. Technical Report 5266. Computer Laboratory, University of Cambridge. British Library Research and Development Report.

K. Sparck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments. Part 1. *Information Processing & Management* 36, 6 (2000), 779–808.

I. Stanton, S. Ieong, and N. Mishra. 2014. Circumlocution in Diagnostic Medical Queries. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 133–142.

P. Thomas, A. Moffat, P. Bailey, and F. Scholer. 2014. Modeling Decision Points in User Search Behavior. In *Proc. IIiX*. 239–242.

P. Thomas, F. Scholer, and A. Moffat. 2013. What Users Do: The Eyes Have It. In *Proc. Asia Information Retrieval Societies Conf. (AIRS)*. 416–427.

E. G. Toms, H. O'Brien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. Macnutt. 2008. Task Effects on Interactive Search: The Query Factor. In *Focused Access to XML Documents*. Springer, 359–372.

A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. 2009. Including Summaries in System Evaluation. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 508–515.

P. Vakkari. 1999. Task Complexity, Problem Structure and Information Actions: Integrating Studies on Information Seeking and Retrieval. *Information Processing & Management* 35, 6 (1999), 819 – 837.

E. M Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing & Management* 36, 5 (2000), 697–716.

E. M. Voorhees. 2002a. Overview of the TREC 2002 Question Answering Track. In *Proc. Text Retrieval Conf. (TREC)*.

E. M. Voorhees. 2002b. Overview of TREC 2002. In *Proc. Text Retrieval Conf. (TREC)*.

E. M. Voorhees. 2003. Overview of the TREC 2003 Robust Retrieval Track. In *Proc. Text Retrieval Conf. (TREC)*.

W. Webber, A. Moffat, and J. Zobel. 2008. Statistical Power in Retrieval Experimentation. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 571–580.

W. Webber, A. Moffat, and J. Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. on Information Systems* 28, 4 (2010), 20.1–20.38.

W. Webber, A. Moffat, J. Zobel, and T. Sakai. 2008. Precision-At-Ten Considered Redundant. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 695–696.

R. W. White and S. M. Drucker. 2007. Investigating behavioral variability in web search. In *Proc. Conf. on the World Wide Web (WWW)*. ACM, 21–30.

R. W. White and D. Kelly. 2006. A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 297–306.

I. H. Witten, A. Moffat, and T. C. Bell. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images* (second ed.). Morgan Kaufmann.

W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. 2012. Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks With Varying Levels of Cognitive Complexity. In *Proc. IIiX*. 254–257.

W.-C. Wu, D. Kelly, and A. Sud. 2014. Using Information Scent and Need for Cognition to Understand Online Search Behavior. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 557–566.

E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. 2010. Expected browsing utility for web search evaluation. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 1561–1564.

E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. 2014. Relevance and Effort: An Analysis of Document Utility. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 91–100.

J. Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments?. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 307–314.

Received 05/2016; revised 09/2016; accepted 12/2016