Towards Nuanced System Evaluation Based on Implicit User Expectations

Paul Thomas[1], Peter Bailey[2], Alistair Moffat[3(✉)], and Falk Scholer[4]

[1] CSIRO, Canberra, Australia
paul.thomas@csiro.au
[2] Microsoft, Canberra, Australia
pbailey@microsoft.com
[3] The University of Melbourne, Melbourne, Australia
ammoffat@unimelb.edu.au
[4] RMIT University, Melbourne, Australia
falk.scholer@rmit.edu.au

Abstract. Information retrieval systems are often evaluated through the use of effectiveness metrics. In the past, the metrics used have corresponded to fixed models of user behavior, presuming, for example, that the user will view a pre-determined number of items in the search engine results page, or that they have a constant probability of advancing from one item in the result page to the next. Recently, a number of proposals for models of user behavior have emerged that are parameterized in terms of the number of relevant documents (or other material) a user expects to be required to address their information need. That recent work has demonstrated that T, the user's *a priori* utility expectation, is correlated with the underlying nature of the information need; and hence that evaluation metrics should be sensitive to T. Here we examine the relationship between the query the user issues, and their anticipated T, seeking syntactic and other clues to guide the subsequent system evaluation. That is, we wish to develop mechanisms that, based on the query alone, can be used to adjust system evaluations so that the experience of the user of the system is better captured in the system's effectiveness score, and hence can be used as a more refined way of comparing systems. This paper reports on a first round of experimentation, and describes the progress (albeit modest) that we have achieved towards that goal.

Keywords: Retrieval evaluation · User behavior · Search user model

1 Introduction

Information retrieval systems underpin the considerable economic success of the web search industry. Billions of queries per day are processed, with search services possessing a seemingly uncanny ability to identify the page or pages that the user is searching for. A key component of retrieval system development is the use of evaluation processes, in order to measure the quality of the results that

G. Zuccon et al. (Eds.): AIRS 2015, LNCS 9460, pp. 332–344, 2015.
DOI: 10.1007/978-3-319-28940-3_26

are returned. Evaluation options include user focus groups, supervised observational trials, unsupervised trials using query and click logs, and corpus-based batch evaluations.

Central to batch evaluation is the notion of an *effectiveness metric*, a mapping from a search engine result ranking to a numeric score. For example, precision at depth k, denoted Prec@k, scores a ranking by the fraction of the first k documents in it that are deemed to be relevant to the query. Many effectiveness metrics have a corresponding *user model*. For example, if the user always examines the first k documents in the ranking, and forms an opinion of the search service according to the number of those k documents that are relevant, then their *expected utility per document inspected* exactly corresponds to Prec@k. A wide range of more gradual weighted-precision metrics – with corresponding user models – have been described. For example, in the *Rank-Biased Precision* (RBP) effectiveness metric [15], the user is assumed to always examine the first document in the ranking, and then, having viewed the document at depth d, go on to depth $d+1$ with a fixed probability p. A range of other metrics follow similar approaches, including *Expected Reciprocal Rank* (ERR) [9].

Our recent work has argued that rather than a fixed probability p, the user begins their search with an implicit goal of fetching T relevant documents and unconsciously adjusts their "continue to the next document" probability as a function of the depth d in the ranking they have reached, and of the extent of their unfulfilled expectation for relevant documents [5,14]. For example, a user who seeks $T = 1$ relevant documents will end their search more quickly than a user seeking $T = 10$ regardless of whether or not relevant documents are encountered; and will end their search even more quickly if they find a high fraction of relevant documents amongst the first ones they examine. These *adaptive* user models should – all other things being held constant – lead to more realistic system comparisons, and hence better-quality outcomes for search users.

There is, however, the question of somehow knowing what the user's expectation of required relevance is – the quantity denoted as T. In a batch evaluation setting, while curating the creation of a test collection, we can simply ask for this quantity directly from the individual at the point of providing a query in response to some information need. In other evaluation settings, interrogating the user ahead of their search activity may not be desirable, let alone possible. As a first step in resolving that uncertainty, the work presented in this paper explores the extent to which T can be established as a function of attributes associated with the query the user issues. Towards that goal, Sect. 2 describes the user models that we work with; Sect. 3 describes a crowdsourced expectation and query data-gathering exercise, and briefly summarizes the results that have already been achieved using that data; and then Sect. 4 describes the additional analysis carried out for the purposes of this work.

Our findings are mixed. Using a range of query-dependent features such as its frequency, and its length, we are able to provide better prediction of T over the aggregated data than simply taking the majority value of T and using it as a constant prediction. However, to date the gain in prediction accuracy that

has been achieved is relatively small – an increase from 29 % to 33 %. While definitely significant, the gain is not substantial, and is rather less than the gain in accuracy resulting from the use of other features including the topic of the query (not normally available in a production system), and the identity of the user that issued the query (which might be available).

2 Background

This section introduces the concepts of search user models, static and adaptive effectiveness metrics, and search task complexity.

User Models: A *search user model* describes the way in which the elements of a ranked list of search results are inspected, and seeks to compute the value of a corresponding effectiveness metric, reflecting the expected rate at which a user gains utility from the system. Gain is a function of relevance, which is assigned to documents by human judges typically using an ordinal scale: for example, a four-level relevance scale might have values *not relevant*, *marginally relevant*, *relevant*, and *highly relevant* [19]. Relevance is then transformed into gain, commonly using a linear [10] or exponential function [8].

Eye tracking analysis has shown that, on average, users scan a search results page from top to bottom [11], although there is substantial additional variation and movement between individual result items [21]. As a result, the search user models behind many evaluation metrics such RBP and NDCG [10] incorporate the notion of a *discount*, where relevant items that are returned lower down a ranked list contribute smaller amounts of utility to the user. Utility gain was originally defined in absolute terms [10]; then in terms of expected utility per document inspected [15]; and most recently, in terms of expected utility per second spent searching [17]. User models where gain is based on rank position alone are called *static*, while those that additionally incorporate information about the relevance of the documents that have been seen earlier in the ranking are called *adaptive*.

Adaptive Effectiveness Metrics: We have recently introduced the notion of the user's expected search goal, quantified as a utility estimate T, hypothesizing that the value of T provides guidance as to the user's behavior while they are scanning the results list [14]. In particular, we suggest that "high T" queries involve the inspection of more documents in the result ranking than do "low T" queries, even before any relevance information is taken in to account. The model is adaptive, with searches in which the anticipated utility is accumulated quickly ending earlier than searches in which relatively few relevant documents are encountered. In followup work, we proposed an effectiveness metric "INST", a weighted-precision "expected utility per document inspected" sum defined by the assumption that the user always examines the first document in the ranking, and then continues from depth i to depth $i+1$ with probability

$$C_{\text{INST}}(i) = \left(\frac{i + T + T_i - 1}{i + T + T_i} \right)^2$$

where T is the user's initial estimate of the number of relevant documents they will find, and T_i is the extent of the relevance found in the first i documents in the ranking. INST brings together a range of desirable attributes in a useful manner: it is adaptive, meaning that for any given value of T, the expected search length is less in rankings with many relevant documents than it is in rankings with few relevant documents; it respects patience, in that the continuation probability $C(i)$ slowly increases towards one, reflecting that a user who has invested heavily in a ranking and is already some way down a list is (on a conditional basis) more likely to continue scanning documents than a user who is still examining documents in the early part of the ranking; and it is not unbounded, since it has a finite expected search depth even on rankings that do not contain any relevant documents at all. As probabilistic limits, the expected search depth on a ranking with no relevant documents is $2T+0.5$, and on a ranking with nothing but relevant documents, is $T + 0.25$ [5].

Search Task Complexity: Users carry out information seeking tasks for many different reasons. A key characteristic that may vary between tasks is their complexity: consider the difference between trying to find the answer to a short factoid question such as the name of the author of "The Odyssey", versus trying to obtain a deeper understanding of the cultural impact of Homer's work. Kelly et al. [12] propose a hierarchy of complexity of search tasks, based on a taxonomy of learning [4]. In the experiments described below, we consider three of Wu et al.'s cognitive complexity levels: *Remember*, tasks that primarily involve factoid-style answers, similar to recalling knowledge from long-term memory; *Understand*, tasks that involve the construction of meaning, for example through interpreting or exemplifying; and *Analyze*, tasks that involve breaking material into parts, and making overall decisions based on how these facets relate to one another [22]. Section 3 gives examples of information needs in these three categories.

Query Variability: Users typically turn to an information retrieval system with the aim of resolving an information need. A key step of their interaction with the system is to translate this information need into a query; for most users of modern search engines, this typically involves typing a small set of search terms into a text box. However, due to the expressiveness of language, many different queries could be used as instantiations of the same information need. This occurs for example when users refine an initial query as part of the same search session [13]. However, attempts to quantify the impact of query variability as a component of IR system evaluation have been limited. As part of the 1999 TREC-8 Query Track [7], participants were asked to generate alternative query strings for supplied information need statements (called search topics in the TREC framework). The track concluded that query variation can lead to substantial differences in retrieval effectiveness. In recent work, we have gathered variant queries intended to express the same underlying information needs [5]. We make use of these crowdsourced queries to investigate approaches to model a searcher's expected utility, as explained in the next section.

3 A Crowdsourced Experiment

As part of our study into the effects of query variability, we carried out a user experiment and gathered data using crowdsourcing. This section provides a brief overview of that experiment, and describes the data that was collected.

Crowdsourcing: Crowdsourcing is the process of soliciting work from a large group of people (the "crowd") in an online setting. The work is typically advertised through a crowdsourcing platform, such as Amazon Mechanical Turk or CrowdFlower. Internet users can register with the platform, search or browse through a list of available work, and choose whether to participate. The terminology of crowdsourcing tends to vary from platform to platform; in this work we refer to the people who offer their labor through a crowdsourcing platform as "workers", and each discrete unit of work that is carried out is called a "task". In the research world, crowdsourcing has become popular as a method to recruit participants for experiments involving humans responses. As an experimental practice, crowdsourcing has been criticized for reducing the level of control that researchers have over their pool of participants; conversely, proponents of crowdsourcing have highlighted that a more diverse user base is likely to be a positive feature, since prior to crowdsourcing the typical participant pool for human factors research studies consisted of university undergraduate students [1]. In the IR field, initial investigations have suggested that crowdsourcing, with appropriate controls to remove "spam" workers (people who do not take the job seriously, or the activities of automated bots intended to mimic human responses), can be a useful source of participants for user studies, including relevance judging [3,18].

Topics and Backstories: The NIST-sponsored TREC shared tasks have been generating useful search data for more than two decades. The test collections (consisting of sets of topics, documents, and judgments) that have been constructed have become invaluable resources for IR experimentation.[1] Table 1 summarizes the three different collections used in our experimentation, and gives a sample "title" query for each collection, noting that detailed "narrative" and "description" sections are also provided for the R03 and T04 topic statements.

In the case of the R03 and T04 queries, we started with TREC topic descriptions and narratives, and wrote what they called a *backstory* for each one, to personalize and motivate the information need. Backstories were also written based on the Q02 questions. For example, the backstories for the queries shown in Table 1 were:

- *You saw a Discovery Channel show that said that it takes eight minutes for the light from the sun to travel to the earth. You want to find out how far away the sun is in miles or kilometers.*
- *A workmate has been diagnosed with arthritis. You know she struggled once with Lyme disease, from a tick bite. You wonder what evidence there is to support (or refute) a connection between the two.*

[1] http://trec.nist.gov.

Table 1. Origins of queries used to create backstories.

Collection	Year	Topics	Example TREC query
Q02	2002	70	1876: how far from the earth is the sun?
R03	2003	60	604: lyme disease arthritis
T04	2004	50	730: gastric bypass complications

– *A surgeon has recently recommended gastric bypass surgery for your overweight uncle. He wants to lose weight, but you would like to help him make an informed decision by alerting him to the possible complications and potential dangers of gastric bypass surgery.*

These three information needs have, respectively, task complexity categories Remember, Analyze, and Understand.[2]

Experimental Process: For each search topic, study participants were first shown the backstory motivating an information need. They were then asked to provide three pieces of information: the total number of useful web pages they thought they would need to look at to answer the information need (T); how many different queries they thought they would need to issue in order to find that number of useful pages (Q); and what their first (written text) query would be when using a search engine to answer the information need. The first two responses were collected using single-selection radio buttons describing numeric ranges, and the third was a free-form text field. The interface, including the range of answer choices, is shown in Fig. 1, using one of the example backstories already introduced. Note the bands on values of T and Q used during the data collection.

The user study was carried out using the CrowdFlower platform.[3] Each task consisted of providing answers in response to five of the 180 different search topics. Users could choose to complete as many units as they wished, providing answers to anywhere between 5 and 180 topics if they wished. Since not all crowdworkers take their tasks seriously, data cleaning was carried out. If any worker entered the same "first query" string for more than one topic, all of their responses were removed from the subsequent analysis. Workers who simply pasted fragments of the topic statements that were deemed nonsensical as their "first query" were also removed. The remaining data consisted of 98 workers who provided 7,969 responses, with a median of 44 responses per topic.

4 Direct Estimates of T and Q

We now describe our detailed investigation of T and Q, making use of the data collected as part of the work described in the previous section.

[2] The backstories are available for reuse at DOI 10.4225/08/55D0B6A098248.
[3] http://www.crowdflower.com.

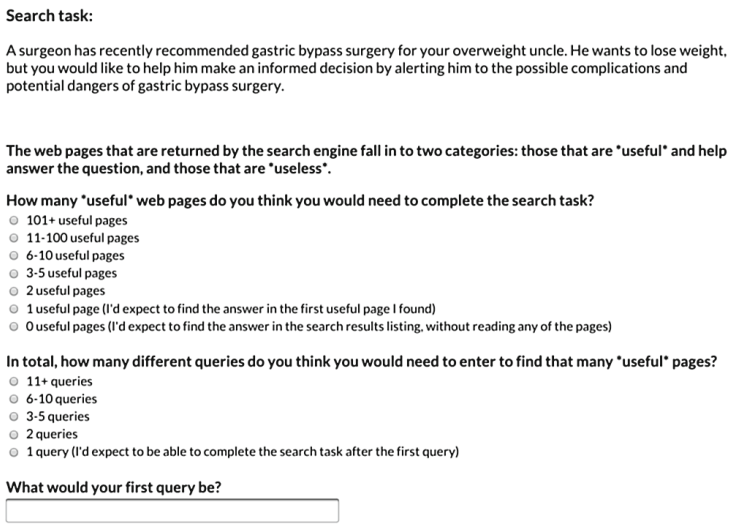

Fig. 1. Screenshot of the CrowdFlower interface.

Overall Distribution: The great majority of responses were for T of one to ten – that is, people expected to read one to ten relevant documents to answer the information need. There were 436 responses (5.5 %) where $T = 0$ and people expected to answer their need from the result listing alone; 654 (8.2 %) of "11–100 useful pages"; and 62 (0.8 %) of "101+ useful pages". The most common responses were $T = 1$, with 2,329 cases (29 %), and $T = 3$–5, with 1,782 (22 %) of responses.

There was a similar skew in estimates of Q. The most common response was "one query", with 3,521 cases (44 %). At the other end of the range, there were only 600 cases (7.5 %) in the top two categories, where workers expected to need six queries or more. Across all 7,969 responses, T and Q are correlated and participants who expect to need several documents also expect to need several queries to find these documents. Figure 2 plots that correlation. As noted in the caption, this relationship is significant according to Spearman's $\rho = 0.66$, one-sided $p \ll 0.01$. A full 24 % of the workers' responses nominated $T = 1$ and $Q = 1$, indicating that they expected to need a single relevant document, and would find it by issuing a single query.

The Influence of Search Task Complexity on T and Q: To investigate the relationship between the complexity of search tasks and the number of documents and queries that a searcher believes they will need to complete the task, each of the 180 search tasks was categorized into one of three search task complexity levels identified by Wu et al. [22].

The relationships between task complexity and the user estimates of the total number of relevant documents that need to be viewed to fulfill an information need (T), and the number of queries that need to be issued to find those doc-

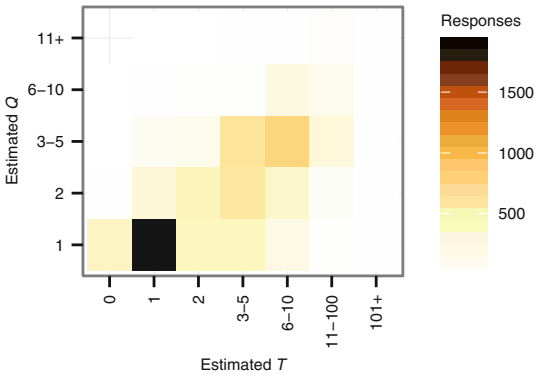

Fig. 2. Correlation between estimates of total number of useful pages (T) and queries (Q) needed to fulfill an information need. The mass along the diagonal bottom-left to top-right demonstrates a positive correlation (Spearman's $\rho = 0.66$, one-sided $p \ll 0.01$).

uments (Q), are depicted in Fig. 3. For both T and Q, a clear trend can be seen whereby a searcher expects to need fewer documents, and fewer queries to find those documents, for tasks of a lower cognitive complexity. The distinction is strongest between the Remember category on the one hand, and the two higher complexity categories on the other; with the Understand and Analyze tasks having somewhat similar distributions. Even so, the highest complexity Analyze category also has the highest overall weight allocated to requiring a larger number of documents and queries.

The demonstration that T and Q are related to the underlying complexity of the information need [5] raises an obvious question: can T and/or Q be estimated or predicted *without* asking the user? To attempt this, and to understand which factors are most relevant, we used cumulative logistic regression[4] and the crowdsourced data to model how T and Q respond to a number of potential explanatory variables. Model selection and parameter estimation were simultaneous, and all data was used to build the model.

– *Per-user and per-topic factors*: We start with two factors which are extrinsic to the query text: the identity of the user (here, CrowdFlower's worker identifier) and the information need or topic (here, the TREC topic number). The first reflects an individual's overall propensity to expect more or fewer interactions, and the second reflects characteristics such as topic complexity. Modern search engines may carry out extensive personalization, and thus can be expected to encode user identity factors relating to long term interaction patterns with documents if they prove useful, within a broader framework [6]. Simpler search engines may not carry out any personalization, and thus would not be able to encode user identity. Even with extensive contextual information modeling, in

[4] Cumulative logistic regression – also known as ordinal regression – used R's `ordinal::clm` and `ordinal::step.clm` functions.

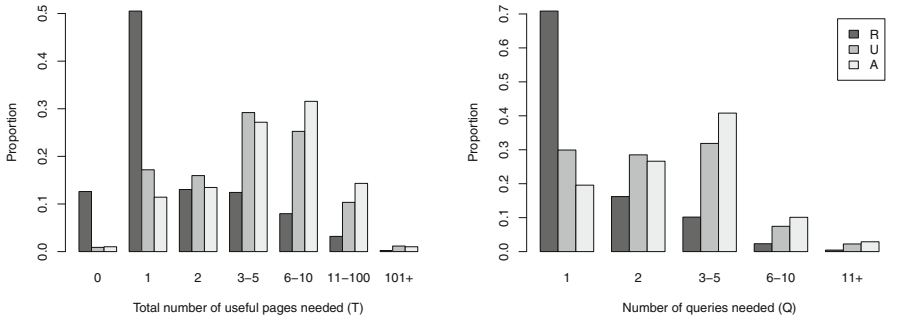

Fig. 3. Total number of useful pages (T) and queries (Q) needed to fulfill an information need of a given complexity level (Remember, Understand, and Analyze).

the general case, an information need (as represented by a topic in our data) is not known to search engines; however topic effects have long been known to be important within test collections. We expect these two factors to explain a lot of variability.

- *Query characteristics*: This includes the number of words and the number of characters – two ways of determining length – and the mean characters per word, which is a surrogate for the complexity of each word. We also investigated the maximum inverse document frequency (IDF) and sum of IDFs assigned by Indri's Okapi BM25 similarity scoring regime, across all terms in the query, as a surrogate for specificity of the query.
- *List characteristics*: Potential explanatory factors include characteristics of the resulting ranked list, which are known after the query is processed but without any further feedback. Here we use the BM25 scores assigned by Indri to the documents ranked 1, 10, and 100, to reflect the quality of the returned list. We also use the ratio of the scores at 10 and 1, 100 and 1, and 100 and 10 (labeled "Indri@10:1", and so on), to reflect the consistency of the scores: a high ratio suggests consistent scores. We also include the number of answers if the query is treated a conjunctive Boolean query – that is, the number of documents containing all query terms.
- *Past behavior*: Finally, we consider two characteristics of past user behavior. Query frequency is drawn from the logs of the Microsoft Bing web search engine, based on usage data within the month June 2015 and is normalized to the range $(0, 1)$, where higher values represent more common queries. Relative click-through rate is drawn from the same logs and represents the ratio of click-through rate for this information need to the global average (so numbers lower than 1 represent fewer clicks). The data is aggregated by topic, and averaged. Missing query-level data is smoothed by assuming the lowest possible query frequency, and the global average click-through rate. This data, of course, is not usually available outside the large search engines.

Fitted Models: Table 2 gives the best models for T and Q, selecting from all of the listed factors except user and topic ID (which may be unknown or unknowable).

Table 2. Significant factors in fitted models for estimates of T and Q. Effect sizes >0 correspond to higher values of T or Q being more likely. All effects significant at $p < 0.05$, Wald test.

Estimating T	
Factor	Effect
Query frequency	–3.46
Mean chars/word	0.15
No. characters	0.03
Indri@100:1	0.96
Max. IDF	–0.07
Relative CTR	0.42
No. words	–0.10
Indri@100:10	–0.86
No. Boolean answers	0.04
Indri@1	0.02

Estimating Q	
Factor	Effect
Query frequency	–4.89
Mean chars/word	0.07
No. characters	0.04
No. words	–0.15
Max. IDF	–0.05
Indri@100	0.01
Relative CTR	0.28

Models were learned to minimize the Akaike information criterion (AIC) [2], which combines log-likelihood with a penalty for each factor in the model. Effects are given as modifiers to log-odds, and effects greater than zero mean higher odds for responses further up the scale – that is, positive effects mean higher values of T are more likely as the underlying variable increases. Factors are given in the order selected. For example, query frequency is the best single factor for predicting either T or Q.

The most predictive factor in both models is query frequency: queries which are more common predict lower values of T or Q. This is easy to interpret, since we may expect popular or authoritative pages for common queries; Teevan et al. [20] have also noted that common queries were less ambiguous. The other behavioral feature, relative click-through rate, has positive and moderately large effects in both models: as click-through rate increases, searchers expect to need more interactions. This means searchers are able to predict, at least crudely, how much interaction they will use to address a need.

Query features are the largest set in both cases. Longer query terms predict more interactions, which is possibly explained by longer terms capturing more complex information needs. Queries with more words predict lower T and Q, which is consistent with longer queries being more specific or possibly asking a question in natural language. This relationship between query length and specificity has been noted by Phan et al. [16], and again Teevan et al. noted queries with more characters were less ambiguous (although this did not hold for queries with more words). As the maximum IDF grows – meaning rarer words appear in the query – we see the same effect, with more specific language corresponding to lower T and Q.

List features are more useful predicting T than they are predicting Q. In the case of T, we see four effects at play. As the number of Boolean answers increases, T increases; T also increases as scores are more consistent from ranks 1 to 100.

Table 3. Summary statistics for trained models for T. Lower ΔAIC, and higher accuracy values, indicate better models.

Model	ΔAIC	Accuracy
Majority	9841	29 %
Query characteristics	6565	32 %
List characteristics	6949	30 %
Past behavior	6875	30 %
Best query-only model	6336	33 %
User only	3870	40 %
Topic only	4907	39 %
User and topic	74	51 %
Best model from all factors	0	51 %

The quality of the first result ("Indri@1") also correlates with T. The consistency of scores in the tail ("Indri@100:10") anticorrelates however, and it is not clear why this is the case. We hope to better understand this relationship in future work.

Models Compared: To further examine how well potential explanatory factors predict T, we built six models on the principles above. The first is a simple majority model (intercept only), and always predicts the most common response, $T = 1$. One model each was built with only query, list, and behavior factors; we also include the learned model reported in Table 2, which draws from all of these sets. Finally, we built a model which uses user and topic identity, which can adapt to per-user preferences and per-topic complexity.

Table 3 reports two measures of quality for each model. Accuracy is the number of times the model exactly predicts our users' estimate of T (recall that users chose from seven bands). ΔAIC is the difference in AIC between each model and the best model we have; lower scores are better. Note that AIC (and hence likelihood) improves dramatically over the majority baseline no matter which factors we use, but query characteristics are the most useful group as a whole with AIC improving by 3176. The combined model is better still with a further AIC improvement of 229. However, accuracy is not significantly better and we only see a 4 % improvement at best, from 29 % to 33 %. If we want to use these models to get a point estimate for T, rather than a distribution over possible values, they are not a great improvement.

If models are allowed to make use of the user's identity – here, we have used the CrowdFlower ID – and the topic behind the query, it is possible to do much better (bottom part of Table 3). Using either of these two factors, a further improvement of over 1600 points of AIC and accuracy of 39–40 % is possible; using both, over 6000 AIC points are gained compared to Table 2, and accuracy of 51 %. If all factors are allowed – that is, user and topic factors as well as query, list, and behavior factors – the best model includes user and topic identifiers, number of words, and mean characters per word, and is better by a further 74 points of AIC while still getting 51 % accuracy.

5 Conclusion

Combined with almost certainly unknowable information (the topic), we were able to achieve an accuracy of 51 % in estimating the users' selection of T from one of seven bands. Using only information from the query and documents, which can be reliably calculated by modern search engines, our best effort achieves only 33 %. Whether this degree of accuracy is sufficient to be useful in practice for evaluation of a search system operating over a query population in-the-wild is unknown. We are forced to conclude that there are other significant factors which we have not considered that contribute to the gap in accuracy for our best performing model.

At least two open issues have been identified with the work as described, which we will address as we continue with this project. First, due to choices made at the time the data was collected, the estimates we have been working with make use of bucketed bands of document and query counts, and involve an inevitable loss of accuracy. We hope to repeat the original experiment with a variation allowing users to provide more fine-grained estimates. Second, our modeling has been attempting to predict T for an individual's estimate relating to an information need. However, the current data actually encompasses a distribution over a number (median 44) of estimates for that information need. Instead of attempting to predict a single estimate, and assuming an information need-centric approach to evaluation (rather than a query-centric approach), we might instead predict a distribution. Adaptive effectiveness metrics such as INST may then require modification to encode T as a probabilistic variable rather than as a fixed value.

In the longer term, we are interested in crafting a fully explicated test collection, starting from information needs and encoding user variability, including queries and effort estimates, across a range of task complexities. We hope that this approach of capturing many sources of variability may assist in closing the gap of modeling effort expectations without explicitly needing to ask the user.

Acknowledgments. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (projects DP110101934 and DP140102655). We thank Xiaolu Lu for assistance with the data collection and Bodo von Billerbeck for assistance with query log mining.

References

1. The roar of the crowd. The Economist (2012)
2. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
3. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In: Proceedings of the SIGIR Workshop. Future IR Evaluation, pp. 15–16 (2009)
4. Anderson, L.W., Krathwohl, D.A.: A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman, New York (2001)

5. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: User variability and IR system evaluation. In: Proceedings of SIGIR, pp. 625–634 (2015)
6. Bennett, P.N., White, R.W., Chu, W., Dumais, S.T., Bailey, P., Borisyuk, F., Cui, X.: Modeling the impact of short-and long-term behavior on search personalization. In: Proceedings of SIGIR, pp. 185–194 (2012)
7. Buckley, C., Walz, J.: The TREC-8 query track. In: Proceedings of TREC 1999. NIST Special Publication 500–246 (1999)
8. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of CIKM, pp. 89–96 (2005)
9. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of CIKM, pp. 621–630 (2009)
10. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002)
11. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of SIGIR, pp. 154–161 (2005)
12. Kelly, D., Arguello, J., Edwards, A., Wu, W.C.: Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In: Proceeding of ICTIR (2015)
13. Lin, S.J., Belkin, N.: Validation of a model of information seeking over multiple search sessions. J. Am. Soc. Inf. Sci. Technol. **56**(4), 393–415 (2005)
14. Moffat, A., Thomas, P., Scholer, F.: Users versus models: what observation tells us about effectiveness metrics. In: Proceedings of CIKM, pp. 659–668 (2013)
15. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. **27**(1), 2:1–2:27 (2008)
16. Phan, N., Bailey, P., Wilkinson, R.: Understanding the relationship of information need specificity to search query length. In: Proceedings of SIGIR, pp. 709–710 (2007)
17. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Proceedings of SIGIR, pp. 95–104 (2012)
18. Smucker, M., Kazai, G., Lease, M.: The TREC-12 crowdsourcing track. In: Proceedings of TREC 2012. NIST Special Publication 500–298 (2012)
19. Sormunen, E.: Liberal relevance criteria of TREC: counting on negligible documents? In: Proceedings of SIGIR, pp. 324–330 (2002)
20. Teevan, J., Dumais, S.T., Liebling, D.J.: To personalize or not to personalize: modeling queries with variation in user intent. In: Proceedings of SIGIR, pp. 163–170 (2008)
21. Thomas, P., Scholer, F., Moffat, A.: What users do: the eyes have it. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 416–427. Springer, Heidelberg (2013)
22. Wu, W.C., Kelly, D., Edwards, A., Arguello, J.: Grannies, tanning beds, tattoos and NASCAR: evaluation of search tasks with varying levels of cognitive complexity. In: Proceedings of IIiX, pp. 254–257 (2012)