Relevance and Effort: An Analysis of Document Utility

Emine Yilmaz, Manisha Verma
Dept. of Computer Science
University College London
{e.yilmaz, m.verma}@cs.ucl.ac.uk

Nick Craswell, Filip Radlinski, Peter Bailey
Microsoft
{nickcr, filiprad, pbailey}@microsoft.com

ABSTRACT

In this paper, we study one important source of the mismatch between user data and relevance judgments, those due to the high degree of effort required by users to identify and consume the information in a document. Information retrieval relevance judges are trained to search for evidence of relevance when assessing documents. For complex documents, this can lead to judges' spending substantial time considering each document. However, in practice, search users are often much more impatient: if they do not see evidence of relevance quickly, they tend to give up.

Relevance judgments sit at the core of test collection construction, and are assumed to model the utility of documents to real users. However, comparisons of judgments with signals of relevance obtained from real users, such as click counts and dwell time, have demonstrated a systematic mismatch.

Our results demonstrate that the amount of effort required to find the relevant information in a document plays an important role in the *utility* of that document to a real user. This effort is ignored in the way relevance judgments are currently obtained, despite the expectation that judges inform us about real users. We propose that if the goal is to evaluate the likelihood of utility to the user, *effort* as well as relevance should be taken into consideration, and possibly characterized independently, when judgments are obtained.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords

Relevance assessments; User behavior; Effort

1. INTRODUCTION

There are two broad approaches to information retrieval effectiveness evaluation: the collection-based approach, and the user-based approach. The collection-based approach measures the effectiveness of retrieval systems using test collections comprising canned information needs and static relevance judgments. These are used to compute evaluation measures such as precision and recall, mean average precision, and many others. On the other hand, the user-based approach involves actual users being observed and measured in terms of their interactions with a retrieval system.

Each of these approaches have strengths and weaknesses. Collection-based evaluations are fast, repeatable, and relatively inexpensive as the data collected can be reused many times. However, collection-based evaluations make many simplifying assumptions about real user information needs, what constitutes relevance, and many other aspects of retrieval (e.g. how summaries are presented to users, etc.). In contrast, user-based approaches can take into account end-to-end user satisfaction. However, they tend to be expensive to perform due to the need to obtain users for every system. They are also difficult to analyze due to the need to control for variance across tasks, population and time. Therefore, collection-based evaluation is commonly used in evaluating the quality of retrieval systems, especially when reusability is a prime concern for enabling rapid experimental iteration among a number of alternatives.

Ideally, the outcome of collection-based evaluation should be predictive of the satisfaction of real users of a search system. Yet research has shown that these two forms of evaluation often do not completely agree with each other [22, 37], or agree with each other only when there is a significant gap in terms of the quality of the systems compared [2, 1]. One of the main reasons behind this mismatch are the simplifying assumptions made in collection-based evaluation about relevance and how users behave when they use a search system. Therefore, there is increasing interest in better modeling user needs and user interaction with an engine in collection-based effectiveness evaluations [11, 10, 9, 36].

We claim that a key source for disagreement between collection-based evaluation and user-based online experiments is due to the disagreements between what judges consider as relevant versus the *utility* of a document with respect to an actual user. The main goal of our work is to identify the reasons behind these disagreements. To address this goal, we rely on implicit feedback (such as dwell time information) provided by users of a retrieval system. We compare indicators of *utility* inferred from implicit feedback to judgments obtained from relevance judges, and identify sources of disagreement.

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661953 .

We further focus on the reasons of mismatch between relevance judgments and implicit signals obtained from the clicked documents (e.g. dwell time) and show that one of the main reasons for such mismatch is due to the effort needed to find and consume relevant information in a document. Relevance judges are explicitly asked to identify the relevance of documents they assess. Therefore, they must evaluate each document thoroughly before marking it as relevant or non-relevant. In performing these judgments, judges often spend a significant amount of effort on documents that may not have significant relevant content or that may be hard to read. On the other hand, users simply wish to fulfill an information need, and are often much less patient when determining if a particular document is relevant. If they do not see evidence of sufficient relevance quickly or if they think relevant information is difficult to consume, they tend to give up and move on to another document.

Therefore, even if a document is relevant to a query, it provides only minimal utility to an actual user if finding and understanding the relevant portions of the document is difficult. Based on this observation, we propose that the judgment process be modified to incorporate how real users interact with documents. This could be done by either explicitly asking the judges to provide information regarding the effort to find relevant information in a document as well as document relevance, or by logging the amount of time it takes a judge to assess the relevance of a document, and using this information in evaluation.

Overall, our findings suggest that *effort* plays an important role in user satisfaction and that further thought should be given to the concept of relevance versus utility. Our results also show that features related to the effort to find/consume relevant information (e.g., readability level of the document, etc.) could be used as ranking features when retrieval systems are designed as they have a significant impact on the utility of a document to an actual user.

In what follows, we start by describing related work in the area. We then follow with a user model that considers the different stages of how users interact with a search engine and compare this with the behavior of judges. Through the user model, we show how the *utility* of a document with respect to an actual user could be different than the relevance of such document. We also show through a regression model that features related to *effort* play a significant role in a relevant document being considered as low utility by an actual user. Therefore, we argue that if we would like to measure the utility of a search system to an actual user, current judging mechanisms should be changed and the assessors should be asked to provide judgments in terms of the *effort* required to find relevant information together with the relevance of the document. We further describe a post-processing heuristic that can be used to infer the utility of a document with respect to an actual user.

2. RELATED WORK

The concept of relevance sits at the core of information retrieval. Hence, much research has been devoted to defining and understanding relevance. Saracevic [33, 32] provides a detailed overview on the definition of relevance and the different approaches to define relevance. Our work supports a pragmatic view of relevance, which is also argued by Saracevic as follows [32]: "It is fine for IR systems to provide relevant information, but the true role is to provide information that has utility – information that helps to directly resolve given problems, that directly bears on given actions, and/or that directly fits into given concerns and interests. Thus, it was argued that relevance is not a proper measure for a true evaluation of IR systems. A true measure should be utilitarian in nature". Cooper [15] was one of the first people who argued for utility instead of relevance as a measure of retrieval effectiveness. He argued that the purpose of retrieval systems should be to retrieve documents that are useful and not solely relevant, and that retrieval systems should be judged based on cost and benefits [32]. In his later paper Saracevic [33] splits relevance into two categories: relevance with respect to the system and user-based relevance. He mentions that effort should be considered when relevance with respect to a user is defined. He further provides a definition of the theory of relevance to an individual and incorporates effort into this definition. Our work proposes and evaluates a specific way to do so.

Recent work has focused on modeling search based on Economic Theory [3, 4] and showed how the search behaviour of users will change as cost and performance change. Through this model, the authors show how cost (or effort) is an important factor that affect user behaviour.

Smucker et al. [36] recently proposed Time Based Gain (TBG) as a new evaluation metric that can incorporate the effort it takes a user to reach a particular document into evaluation. This work is a step forward in the direction of considering effort spent by a user in collection-based evaluation of retrieval systems. However, TBG does not really focus on the effort needed to find relevant information in a particular document and the effect of this on the satifaction of users from reading that document. TBG mainly assumes that the discount functions used in the metrics should be a function of time it takes a user to reach a particular document (rather than the position of the document) and does not question the definition of *relevance* or incorporate *effort* into the concept of *relevance*. In work motivated by XML and multimedia retrieval systems, de Vries et al. [16] present a user model based around a "tolerance to irrelevance". They propose a model where users start reading from some document entry point and continue until either satisfied with relevant information and/or that relevant information "is starting to appear", or reach their time-based (or user effort-based) irrelevance threshold. Upon reaching an irrelevance threshold, the users proceed to the next system result. This work shares similarities with ours in terms of an abstract user model, but was motivated more by addressing the issues of not having a predefined retrieval unit within video and XML retrieval test collections, whereas ours seeks to improve the relevance judgment collection process by capturing true user behavior (akin to the "tolerance to irrelevance" concept) more successfully. Separate line of research has focused on effort needed to judge the relevance of a document. This work showed that relevant documents require more effort to judge and that effort increases as document size increases [38].

Most recent work on relevance has focused on 1) how to measure the quality of relevance judgments obtained from experts and how evaluation is affected by the quality of relevance judgments, and 2) inferring relevance and evaluating quality of retrieval systems using implicit user behavior. Below we summarize the related work in these areas.

Measuring Relevance Judgment Quality

In the field of information retrieval, assessor agreement has been one of the main criteria used in evaluating the quality of relevance judgments used to in a test collection. Recently, Scholer et al. [34] studied disagreements between judgments in test collections by identifying judged duplicate documents. They found a high level of inconsistency among judgments in a number of standard collections, and analyzed when and why these occur. A significant amount of similar work has been devoted to analyzing the reasons for disagreements in relevance judgments and how these disagreements may affect retrieval evaluation. Further reasons for disagreement between different relevance assessors, such as the instructions given to judges or the different topics have also been analyzed by Webber and colleagues [41, 42]. Also, Chandar et al. [13] analyzed how features such as the reading level of a document and document length affect assessor disagreement and showed that reading level features are negatively correlated with disagreement. They also showed that shorter documents that are easier to understand provoke higher disagreement and that there is a weak relationship between document length and disagreement between the judges. This body of work shows that even in highly controlled judging scenarios, it is difficult for experts to reliably assess what constitutes relevance in documents.

Substantial work has also been devoted to analyzing error in judgments and their effect on evaluating the quality of retrieval systems, as opposed to just focusing on the disagreements in relevance judgments. In older work, Voorhees et al. [39] showed that even though judges may disagree as to what is considered as relevant, these disagreements do not affect the relative ranking of information retrieval systems in TREC. Conversely, Bailey et al. [5] found that degrees of task-expertise and topic-expertise in judges could affect relative ranking of systems within the TREC Enterprise track results. Carterette et al. [12] also showed that systematic judgment errors can have a large effect on system rankings especially when the test collections contain a large number of queries with limited relevance judgments. Meanwhile, Kazai et al. [24] focused on systematic judgment errors in information retrieval and showed that systematic differences in judgments are correlated with specific characteristics of the queries and URLs.

Inferring Relevance from Usage

Using clickthrough data to infer the relevance of documents in online information retrieval systems has also attracted a lot of attention in the information retrieval community [23, 18, 17]. It has been found that clickthrough statistics are often highly affected by issues such as presentation bias and perceived relevance of the documents. The perceived relevance of a document is mainly based on the summary (snippet) of the document and can be different than the actual relevance of the document; hence, users may end up clicking on a document and find out that it is not relevant [17].

In order to overcome this problem, dwell time, the time spent examining a document, has been proposed as an implicit signal of relevance and dwell time is shown to be a good indicator of user satisfaction [21, 19, 25, 8, 43, 14]. There has been many different studies comparing dwell time with relevance. Kelly et al. [26] gives an overview of different research that has been done to analyse the correlations between dwell time and relevance [27]. The Curious Browser experiments showed that when users spend very little time on a page and go back to the results list, they are very likely to be dissatisfied by the results, and that a dwell time threshold of 20/30 seconds could be used to predict user satisfaction [20]. Morita and Shinoda [30] examined the relationship of three variables on reading time: the length of the document, the readability of the document and the number of news items waiting to be read in the user's news queue. Very low correlations (not significant) were found between the length of the article and reading time, the readability of an article and reading time and the size of the user's news queue and reading time. Based on these results, the authors examined several reading time thresholds for identifying interesting documents. When applied to their data set, they found that the most effective threshold was 20 seconds, resulting in 30% of interesting articles being identified at 70% precision. Later, Buscher et al. [8] showed that using documents with dwell time less than 30 seconds as negative feedback resulted in better improvements in ranking performance than any other dwell time thresholds. They further showed that showing users only documents that have dwell time greater than 30 seconds have resulted in the best average precision and NDCG scores. Over many years, a dwell time value of 30 seconds has become the standard threshold used to predict user satisfaction [35, 21, 8, 40, 21, 28]. In general, a very low dwell time can be reliably used to identify irrelevant documents. The converse of this is not necessarily true: a user may spend a long time searching for relevant information in a document and may fail to find the needed information. Hence, long dwell does not necessarily imply relevance [21]. Furthermore, the dwell time threshold that can be used to predict relevance is shown to vary depending on the task [28, 25]. Therefore, it is difficult to say that a document with dwell time above a certain threshold is relevant, whereas a document with a very low dwell time is likely nonrelevant [21].

As implicit signals such as dwell time provide insight about actual users, a substantial amount of work has also looked at how judgment-based metrics can be improved using usage behavior. Most of these focus on devising evaluation metrics that combine document-based relevance judgments in a way that better models user behavior [36, 11, 10, 9].

3. USER BEHAVIOR AND RELEVANCE

As our focus is on the difference between the relevance judgments obtained from judges and the utility of a document to a real user, our first step is to consider how users assess documents. We now present a user model for how actual users behave *after* they click on a document when they use a search engine, given a real information need. We compare this model to the model that is implicit in current relevance assessment guidelines.

3.1 User Model

We propose a simple two-stage model of how real users behave when considering a search result returned by an information retrieval system such as a web search engine:

- **Stage 1: *Initial Assessment***
  Upon clicking on a document (with an expectation of finding relevant content), users make a rapid adjustment to their expectation. They focus on questions such as "Can I find value here?", "How much time do

I need to spend to find the information I need in this document?", "Can I understand this document?" and "Can I find what part of the document is relevant?".

- **Stage 2:** ***Extract Utility***
  Assuming the user expects that they can extract value by identifying an answer to their question or information need, the user is now willing to commit time to read long-form content, view multimedia, or complete a transaction.

After clicking on a document, a *real user* needs to go through both stages to extract value. However, if a document does not seem promising during the initial stage-one assessment, the user may give up: The user has decided it is not worth the risk and effort to continue, and this may be a rational decision even if the document is relevant. A better investment of effort could be to try another document, try another query or give up entirely on the search. Trying for another document may be a particularly good strategy if the user expects that another relevant document exists that is much simpler to consume.

On the other hand, *relevance judges* mainly go through Stage 1: their goal is only to identify whether a document is relevant. They do not need to then consume that information fully. As accuracy in judgments is the key criteria in assessing judgment quality, judges are more willing to invest time to ensure that their answer to the Stage 1 assessment is correct. Judges also sometimes spend substantial time deciding the degree of relevance, for instance considering guidelines to determine if a document should be marked as relevant or highly relevant.

We conclude that for documents where judges take a long time before making a relevance assessment, this assessment itself is difficult. The judges spend all this time in the first stage. Therefore, in this paper we take long judging time to indicate a *high-effort* document. In contrast, we hypothesize that when users spend a long time on a document, they are either spending time consuming the content (stage 2) or this is a case where even users are willing to spend a long time on Stage 1.

To summarize, when the dwell time (time spent on a document by actual users) and the judgment time spent on a document are considered with respect to the above user model, one of the following four cases must hold for the reasons detailed below:

1. Low dwell time, low judgment time.

One possibility in this case is that the document is obviously non-relevant, and both users and judges reach this assessment quickly. Alternatively, the document may be relevant for the information need such that the second stage can be completed quickly. For instance, a question-answering information need may require users to simply read a single sentence or number to extract utility.

2. High dwell time, low judgment time.

This scenario would occur when a document is clearly relevant, and real users spend substantial time on the second stage extracting utility from the document.

3. Low dwell time, high judgment time.

As judges take a long time, it is unlikely that the document is obviously relevant or obviously non-relevant. This suggests that users are abandoning the document in Stage 1 because it does not appear promising. The document could appear non-relevant at first, or the relevant portion could be difficult to find. The document might contain too many advertisements, unnecessary images or other obfuscating content, that make it difficult to consume or might make users doubt the reliability of the content. Regardless of relevance, a document that users tend to give up on is a low value (utility) search result for those users. Hence the question of relevance is moot.

4. High dwell time, high judgment time.

This is a document returned for a task where the user has an important information need that can not be easily answered, perhaps requiring some in-depth consideration of the document's content. Because it is important to them, the user is willing to put in the time in Stage 1 to find the answer. The document could be relevant or non-relevant. If the document is non-relevant, it is a particularly harmful document, because it is requiring high effort from users, and does not pay off. However, a judge can correctly judge such a document to be relevant, and such a document is of value to users.

Next, we describe how we use this user model to infer the utility of documents with respect to actual users and show that of the four cases, most mismatches between relevance versus utility of a document tend to occur on documents under case 3, documents that the users do not spend much time on but are labeled as relevant. We further show that these mismatches are mainly caused by effort to find relevant information.

4. EXPERIMENTAL SETUP

In order to compare relevance judgments obtained from judges with the usefulness of the document to actual users, we use three datasets. Each dataset is parameterized by three aspects: (a) the source of queries that are judged, (b) the types of judges performing the judgments, and (c) the way in which dwell time data was collected.

As our first and second datasets we use data from Kazai et al [24], which was used to analyze systematic judging errors in information retrieval. This data consists of queries from TREC Web Track Ad Hoc task in 2009 and 2010. It was constructed by scraping the top 10 search results from Google and Bing for the 100 queries from Web Track 2009 and 2010, resulting in a total of 1603 unique query-URL pairs over the 100 topics. This method of re-sampling documents for the TREC topics was preferred in order to ensure up to date coverage of the topics and high overlap with the query-document pairs that appear in the logs of the commercial search engine, which we aim to use in our analysis. These 1603 query-URL pairs were then judged by highly trained judges (experts) that are employed by a commercial search engine, as well as crowdsourced judges, forming two different datasets, the ExpertJ-TrecQ and the CrowdJ-TrecQ datasets. To be comparible with judgments from TREC Web Track, each query-document pair in these dataset are labelled on a five-point scale from *bad* to *perfect* and the majority vote was used to identify the final label for a document. More details on the judging setup can be found at Kazai et al. [24].

Since our goal is to study utility of a document with respect to an actual user versus a relevance assessor, we sim-

Table 1: Datasets used for analysis

Dataset	Queries	Judges	Clicks collected on
CrowdJ-TrecQ	Manually constructed for TREC	Crowdsourced	Natural search rankings
ExpertJ-TrecQ	Manually constructed for TREC	Trained experts	Natural search rankings
CrowdJ-NaturalQ	Sampled from actual query distribution	Crowdsourced	Randomized rankings

plify the analysis by only considering relevance as a binary notion, converting the graded relevance judgments into binary. In our analysis, all documents labelled as bad are taken as non-relevant, and all others are taken as relevant to the query. We have tried alterhative ways of obtaining binary judgments from multi graded judgments (e.g. assuming grades 0 and 1 are non-relavant and the rest are relevant, etc.). However, the conclusions of the paper were not affected by how binary judgments were

Our third dataset (CrowdJ-NaturalQ) consists of queries sampled from the actual traffic of a commercial search engine. We mined the anonymized query logs from a commercial search engine for a seven-week period starting in late 2012 and extracted queries and result clicks for further analysis. To reduce variability from cultural and linguistic variations in search behavior, we only included log entries from searchers in the English-speaking United States locale. We restrict our analysis to pairs of URLs shown in the top two positions of the organic Web results, where for the same query the URLs in the pair would exchange ordering to minimize the presentation bias, based on the FairPairs algorithm [31].

In this dataset, each document was labelled as relevant or non-relevant by five judges via crowdsourcing and the label that gets the majority vote is assigned as the final relevance label for the document.

We further restrict our analysis to those pairs for which each ordering has a minimum of 30 impressions (since not all impressions had clicks), and we then filter to pairs of URLs where the clickthrough rate (CTR = number of clicks / impressions) for one ordering is significantly greater than it is for the other ordering with $p < 0.05$. Finally, of the pairs with significantly different CTR as just determined, we take a stratified sample, which is stratified both along the query frequency dimension and the CTR dimension to produce a set of about 5,000 query-document-document tripes.

We use the same judging interface as the one used for the CrowdJ-TrecQ judges, where each document is labelled on a five point scale from *bad* to *perfect* by five judges via crowdsourcing and the label that gets the majority vote is assigned as the final relevance label for the document. The properties of all three datasets are summarized in Table 1.

Measuring Dwell Time: To get the dwell time information for the documents in our datasets, we use click logs of a commercial search engine over a 3 month period starting in September 2013. The dwell time information was collected by observing all clicks on the search engine results during this period. Note that dwell time information was not available for all the documents judged in the datasets as users tend to click only on documents that they assume will be relevant based on the document snippet. To make sure that we have a reliable estimate for dwell time, we focus on documents that have been clicked at least 30 times during the 3 month period. When the documents for which we have dwell time information are available are considered, we end up with 4399 documents for the CrowdJ-NaturalQ dataset, and 1538 documents for the ExpertJ-TrecQ and CrowdJ-TrecQ datasets. For each document, we have dwell time information from many different users and we use the median dwell time on a document as the dwell time for that document as median dwell time was shown to be a more reliable indicator of relevance than the mean [43].

Since our datasets are constructed by using frequently-clicked documents for which reliable dwell information is available, a significant proportion of documents are labeled as relevant by our judges. This does not constitute a problem for our analysis as we are mainly interested in studying the documents that are labelled as relevant but are of low utility to the users.

Measuring Judgment Time: Each query-document pair in all our three datasets is labeled by multiple judges. We use the median judging time spent across all judges as the judging time needed to label a particular document.

5. EXPERIMENTAL RESULTS

The left-side plot in Figure 1 shows the cumulative distribution for the judging time for ExpertJ-TrecQ and CrowdJ-TrecQ datasets versus the dwell time on these datasets. The plot shows that expert judges usually require more time to label a documents than crowd judges: 95% of the documents were labeled within 140 seconds by the expert judges as opposed to approximately 90 seconds for the crowd judges. The plot also shows that on certain documents users spend substantially longer time than the judges. This is expected according to our user model as users tend to go through both Stage 1 and Stage 2 of the user model if they decide that the document is worth examining in Stage 1 whereas the judges mainly go through stage 1. The right figure in the plot shows how dwell time on a document compares with judging time on that document for the CrowdJ-TrecQ Dataset. It can be seen that there is no linear correlation between dwell time and judge time – judges may spend long time judging documents that have a low dwell time and vice versa. The other datasets have very similar behavior to the CrowdJ-TrecQ dataset; hence the plots corresponding to the other datasets are omitted due to space limitations.

5.1 Utility versus Relevance

Our main focus in this paper is to study the *utility* of a document with respect to an actual user versus the relevance of the document. For this purpose, we divide our three datasets into the four scenarios that might happen according to our user model in Section 3 by computing the number of relevant documents versus total number of documents that have 1) low dwell time and low judgment time, 2) high dwell time and low judgment time, 3) low dwell time and high judgment time, and 4) high dwell time and high judgment time. In order to identify documents with low/high dwell time and low/high judgment time, we use the following strategies:

Low vs. High Dwell Time: In Section 2 we provide an overview of how dwell time has been used in the literature

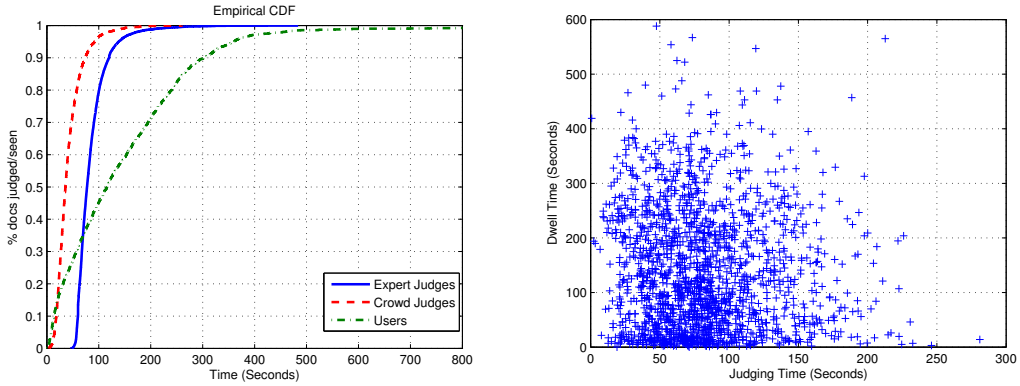

Figure 1: (Left) Cumulative distribution of judgment time for crowd and expert judges versus dwell time, and (Right) Judging time versus dwell time.

Table 2: Dwell Time vs. Judging Time for CrowdJ-TrecQ Dataset

	High Dwell	Low Dwell
High Judg. Time	593/625	**112/134**
Low Judg. Time	650/654	116/125

Table 3: Dwell Time vs. Judging Time for ExpertJ-TrecQ Dataset

	High Dwell	Low Dwell
High Judg. Time	588/644	**88/114**
Low Judg. Time	595/635	121/145

to infer the relevance of a document. Previous work showed that a short dwell time (typically less than 20 or 30 seconds) reliably indicates that the document was not found to be relevant by the user, as he or she decided to stop considering the document quickly [21, 19, 25, 8, 43]. More recently a dwell time threshold of 30 seconds has become the standard threshold used to predict user satisfaction [35, 21, 8, 40, 21, 28]. Therefore, we use a dwell time threshold of 30 seconds to identify documents with low dwell time versus documents with high dwell time. We have also repeated our experiments using a dwell time threshold of 20 seconds, which did not lead to any differences regarding the conclusions of the paper. Therefore, we mainly report results obtained using a dwell time threshold of 30 seconds.

Low vs. High Judgment Time: For each dataset, we use the median judgment time in that dataset as the threshold to identify documents with low vs. high judgment time.

Given these definitions of low vs. high dwell time and low vs. high judgment time, Table 2, Table 3, and Table 4 show the total number of relevant documents versus the total number of documents for each of the four cases of the user model. Considering each of the four cases separately with respect to the user model enables us to analyse the utility of a document with respect to an actual user versus relevance of a document.

1. Low dwell time, low judgment time: According to our user model, these documents can be either relevant or non-relevant: if it is non-relevant, both users and judges reach this assessment very quickly and if it is relevant, the relevant information could be identified and consumed quickly. Given these, documents that fall under this category could be of high or low utility to an actual user. In our datasets, most documents that fall under this category tend to be labeled as relevant by the judges (116 out of 125 documents for the CrowdJ-TrecQ dataset, 121 out of 145 documents for the ExpertJ-TrecQ dataset, and 1974 out of 1987 documents for the CrowdJ-NaturalQ dataset), which is reasonable given that our dataset mainly consists of documents with many clicks.

Table 4: Dwell Time vs. Judging Time for CrowdJ-NaturalQ Dataset

	High Dwell	Low Dwell
High Judg. Time	1903/1957	**213/218**
Low Judg. Time	1974/1987)	236/237

2. High dwell time, low judgment time According to our user model this scenario would occur when a document is clearly relevant, and real users spend substantial time on the second stage extracting utility from the document. Therefore, almost all documents that fall into this category are marked as relevant by our judges, which is expected. However, it is difficult to say whether these documents were of high utility to the user in the end as the user may still not be able to extract the required information from the document even after spending a long time on it.

3. Low dwell time, high judgment time As judges take a long time, it is unlikely that these document are obviously relevant or non-relevant. However, given the low dwell time on these documents, the users tend not to spend the effort to understand whether the document is relevant. The document could appear non-relevant at first, or the relevant portion could be difficult to find or consume. Therefore, documents that fall under this category tend to be of low utility to the users even if they may have relevant content. However, most of these documents are still labeled as relevant by our judges (112 out of 134 (84%)for the CrowdJ-TrecQ dataset, 88 out of 114 (77%) for the ExpertJ-TrecQ dataset, and 213 out of 218 (98%) for the CrowdJ-NaturalQ dataset).

4. High dwell time, high judgment time This is a document returned for a task where the user has an important information need that can not be easily answered, perhaps requiring some in-depth consideration of the document's content. Because it is important to them, the user is willing to put in the time in Stage 1 to find the answer. The document could be of high or low utility to the user, as

in the end the user may decide that the document does not contain the information they need. However, a judge can usually correctly judge the relevance of such a document, as well as its utility to users.

Out of these 4 cases, we are mainly interested in case (3) as in that case the utility of a document to an actual user seems to be different than the relevance of the document. Our hypothesis in this paper is that this possible difference between utility and relevance could be occurring due to a judge spending too much time on a document that a user would not be willing to spend (case 3 in Section 3.1).

Figure 2 shows how the percentage of documents that are of low utility with respect to the users but are labeled as relevant changes as the difference between judge time and dwell time varies for the (left plot) CrowdJ-TrecQ , (middle plot) ExpertJ-TrecQ , and (right plot) CrowdJ-NaturalQ datasets. This figure was generated by calculating the difference between judgment time and dwell time, then grouping documents into buckets of length 20 seconds according to this difference. It can be seen that as the difference between judge time and dwell time increases, the percentage of documents that are labeled as relevant but of low utility to users tend to also increase, showing that such cases are more likely to happen when the judges spend significantly more time on a document than the users.

Judges are likely to spend more time on documents that require a high effort to find and extract relevant information and the users quite often may not be willing to put in this effort. Therefore, our hypothesis is that the mismatches between utility and relevance are likely to be caused by factors related to effort. In the next section, we show that effort is indeed a significant factor that causes the disagreements between relevance versus utility of a document.

5.2 Effect of Effort on Utility versus Relevance

Given that under case 3 of our user model most documents that are of low utility to users are labeled as relevant in our dataset, we next analyze the factors that might cause these disagreements between utility and relevance. As shown in the previous section, most of these mismatches tend to occur when judges spend a long time judging documents that users quickly decide to be of low utility. Our further hypothesis is that these may be *high-effort* documents, where people need to work relatively hard to extract relevant information, and users decide it is not worth the effort. This might be due to the document being too long, document being too difficult to read or other factors.

In order to validate this hypothesis, we extracted several features that might be related to the effort required to read a document. Table 5 provides a list of the features we have used for this purpose. In particular, we focused on features related to the readability level of the document, the document length and the location of the query terms in the document.

Recent work has shown that users tend to scan the documents and only read parts of the document they find relevant [7]. Hence, we also we do not assume that the user reads the entire document, but instead they may search for query terms and read sentences where the terms appear. Therefore, apart from features related to the readability level of the entire document, we also extracted features related to the readability of the sentences in which query terms occur. We assume that users who find the query term in a sentence tend to also read the previous and next sentences and we also include these sentences when the features for the sentences with query terms are extracted.

Table 5: Features related to effort in reading a document.

Name	Description
ARI(d_i)	Automated Readability Index of d_i
LIX(d_i)	LIX Index of d_i
numsent(d_i)	Number of sentences in d_i
numwords(d_i)	Number of words in d_i
ARI($sentquery_i$)	Automated Readability Index of sentences with query terms in d_i
LIX($sentquery_i$)	LIX Index of of sentences with query terms in d_i
numsent($sentquery_i$)	Number of sentences with query terms in d_i
numwords($sentquery_i$)	Number of words in sentences with query terms in d_i
numQ($sentquery_i$)	Number of query terms in sentences with query terms in d_i

To measure the readability of the documents and sentences with query terms, similar to the statistics used by Chandar et al. [13] to analyze the effect of the readability level of a document on assessor disagreement, we use Automated Readability Index (ARI) and LIX. Automated Readability Index (ARI) [29] produces an approximate representation of the US grade level needed to understand the text and is defined as follows:

$$ARI = 4.7\frac{chars(d_i)}{words(d_i)} + 0.5\frac{words(d_i)}{sent(d_i)} - 21.43 \qquad (1)$$

where $char(d_i)$ is the number of letters, numbers, and punctuation marks, $words(d_i)$ is the number of words, and $sent(d_i)$ is the number of sentences.

LIX [6] is another index used to represent readability level of a document. It can be computed as :

$$LIX = \frac{words(d_i)}{period(d_i)} + \frac{longWords(d_i) * 100}{words(d_i)} \qquad (2)$$

where $words(d_i)$ is the number of words, $period(d_i)$ is the number of periods (defined by period, colon or capital first letter), and $longWords(d_i)$ is the number of long words (more than 6 letters) in a document.

These readability estimates output a grade level, where higher level indicates that more effort is required to read the document. Readability level of sentences with query terms are calculated by treating them as independent documents. Primary features for the document (d_i) and sentences in d_i that contain the query terms ($sentquery_i$) are summarized in the Table 5.

We first analyze the factors that might cause users to spend different amounts of times on clicked documents. Therefore, we build a regression model that predicts the dwell time from the features we extracted and identify the factors that have a significant contribution for predicting dwell time.

Table 6 shows the features we used in our regression model and the corresponding p values for these features. The features that have a significant contribution to the model

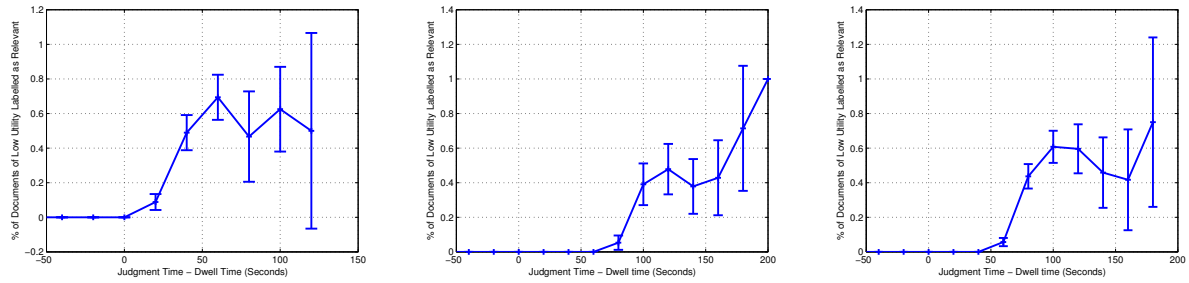

Figure 2: Percentage of low utility documents labeled as relevant versus difference between judging time and dwell time for the (left) CrowdJ-TrecQ , (middlle) ExpertJ-TrecQ , and (right) CrowdJ-NaturalQ datasets.

with $p <= 0.05$ are highlighted in bold. Note that since the ExpertJ-TrecQ and CrowdJ-TrecQ datasets contain the same query-document pairs, there is a single column in the table labeled $TRECQ$ for these two datasets, and another column for the NaturalQ dataset. It can be seen that features related to the readability of the document and the sentences with query terms ($ARI(d_i)$ and $ARI(sentquery_i)$) seem to have a significant effect on the amount of time users spend on the documents on the $TRECQ$ dataset. For the CrowdJ-NaturalQ dataset, readability of the entire document does not seem as important as the readability of the sentences with query terms, as $LIX(sentquery_i)$ and $numwords(sentquery_i)$ seem to be significant factors in the model whereas $LIX(d_i)$ / $ARI(d_i)$ or $numwords(d_i)$ do not seem to be significant. We believe that this is due to the properties of the dataset: queries in the CrowdJ-NaturalQ dataset are sampled from the logs of a real search engine where most queries tend to be navigational and it is easy to spot the parts of the documents that are relevant to the information needs for these types of queries.

Of the significant regression factors, some have positive regression weights and some negative. The weight associated with readability of the entire document, $ARI(d_i)$, is positive for the $TRECQ$ data. This means that users tend to spend longer time on documents that are more difficult to consume. From our current data we can not tell whether they are spending the additional time on Stage 1 or Stage 2 of our user model (Section 3.1). Meanwhile, weights associated with sentence-level readability, such as $ARI(sentquery_i)$, are negative. Similarly, in the case of $NaturalQ$ dataset, factors that are related to the readability level of the sentences with query terms, such as $LIX(sentquery_i)$ and the length of the sentences with query terms ($numwords(sentquery_i)$) tend to get negative waits, suggesting that they have a negative effect on the total amount of time users spend on these documents. This reduction in dwell time when relevant sections are difficult to consume seems to be due to users searching for the query terms in the document and giving up quickly when they realize that these sections are too difficult to read.

Given that the readability level of a document has a significant impact on the amount of time actual users spend on the document, we then focus on the reasons as to why the utility of a document with respect to actual users sometimes differ from the relevance of a document (Upper right cells in Table 2, Table 3 and Table 4). Therefore, we build a regression model that predicts whether the utility of a document will be different than the relevance of a document (case (3) of our user model), given that users spend a short time on them (case(1) and case(3) of our user model). Table 7 shows the p values for the different features we used in our model. Features that are related to the readability of the entire document , such as $LIX(d_i)$, $numsent(d_i)$ and $numwords(d_i)$ seem to have an important contribution as to why users find a relevant document of low utility. All the significant features have a positive contribution to the model, validating our hypothesis that the effort required to find relevant information in a document is highly important for users whereas is is not considered by relevance assessors, suggesting that if we are interested in measuring the utility of a document to an actual user, we should incorporate effort as well as relevance into our judging procedure.

Table 6: Significance of features for predicting dwell time.

Feature	p value	
	TRECQ	NaturalQ
ARI(d_i)	**0.001**$^+$	0.587
LIX(d_i)	0.409	0.103
numsent(d_i)	0.935	0.539
numwords(d_i)	0.718	0.175
ARI($sentquery_i$)	**0.004**$^-$	0.517
LIX($sentquery_i$)	0.562	**0.001** $^-$
numsent($sentquery_i$)	0.588	0.729
numwords($sentquery_i$)	0.600	**0.026** $^-$
numQ($sentquery_i$)	0.504	0.219

Table 7: Significance of features for predicting the mismatch between utility and relevance.

Feature	p value		
	ExpertJ-TRECQ	CrowdJ-TRECQ	CrowdJ-NaturalQ
ARI(d_i)	0.067	0.636	0.710
LIX(d_i)	**0.050**$^+$	0.791	**0.005**$^+$
numsent(d_i	0.138	**0.032**$^+$	**0.038**$^+$
numwords(d_i)	**0.045**$^+$	**0.009**$^+$	**0.050**$^+$
ARI($sentquery_i$)	0.696	0.694	0.333
LIX($sentquery_i$)	0.078	0.221	**0.000**$^+$
queryTerms($sentquery_i$)	0.44	0.478	0.708
numsent($sentquery_i$)	0.386	0.280	0.063
numwords($sentquery_i$)	0.271	0.236	0.273
numQ($sentquery_i$)	0.401	0.112	0.348

Table 8: Dwell Time vs. Judging Time for CrowdJ-TrecQ Dataset

	High Dwell	Low Dwell
High Judg. Time	593/625	**1/134**
Low Judg. Time	650/654	114/125

Table 9: Dwell Time vs. Judging Time for ExpertJ-TrecQ Dataset

	High Dwell	Low Dwell
High Judg. Time	529/644	**0/114**
Low Judg. Time	595/635	101/145

5.3 Combining Effort and Relevance

Our results in the previous sections validate our hypothesis that effort required to find/consume relevant information is an important factor in the utility of a document to an actual user. This observation suggests that the guidelines given to relevance assessors should be changed and the assessors should provide judgments in terms of the effort required to find/consume relevant information in a document, as well as the relevance of the document. Another possibility could be to log the time judges tend to spend on documents and take this time into consideration together with relevance when the utility of a system with respect to actual users is computed.

For our three datasets, we do not have the judgments related to effort. Therefore, we will be simulating how judgments on effort and relevance could be combined to infer the utility of a document by using dwell time as an approximation to effort users are willing to put into reading a document. As mentioned before, real users usually go through both stages of the user model in Section 3.1 whereas judges tend to only go through the Stage 1 of the user model. Hence, dwell time gives us the total amount of time a typical user would spend on the the first and second stages of the model. Therefore, dwell time of a document could be a reasonable upper bound for how long a typical user would be willing to spend on the first stage. If the judge takes significantly more time than that, it is likely that the amount of time needed to find relevant information in the document is more than what most users are willing to spend, therefore this document of low utility to the user as it requires high effort.

Based on this observation, we apply a simple heuristic in which we limit the judgment time on a document to median dwell time on that document. We assume that any document that takes more time to judge than the dwell time is of low utility (even if it might be labeled as relevant). Table 8, Table 9, and Table 10 shows the number of documents that are of high utility to the users (relevant and most users are willing to invest the effort) versus the total number of documents for each of the four categories according to our user model. According to our user model, the documents that fall under low dwell/high judgment time category are the ones that are likely to be of low utility to users. It can be seen that by using this heuristic, we can correctly infer the utility of such documents even though they might be judged relevant. For example, out of the 134 documents that fall under this category for the CrowdJ-TrecQ Dataset, 112 of them were labeled as relevant by our relevance assessors (Table 2). By applying our heuristic for inferring utility, we can infer that almost all these documents are of low utility to the users.

Table 10: Dwell Time vs. Judging Time for CrowdJ-NaturalQ Dataset

	High Dwell	Low Dwell
High Judg. Time	1801/1957	**3/218**
Low Judg. Time	1974/1987	231/237

Applying such a heuristic also results in inferring that some relevant documents with high dwell time and high judgment time are of low utility to the users. For example, out of the 635 such documents for the ExpertJ-TrecQ Dataset, 588 of them were labeled as relevant (Table 3) but by applying such a heuristic, we infer that only 529 of such documents are of high utility to the users, which may not necessarily be correct. However, the number of such cases is not significant and overall the numbers in Table 8, Table 9, and Table 10 look much closer to the expected utility of the documents with respect to the four categories of our user model when compared to Table 2, Table 3, and Table 4.

It is important to note that this correction heuristic was performed without the judges' knowledge. It is likely that some of the long judgment times were the result of judges studying the documents to assess the exact level of relevance, having established that there is relevant content in the document quickly. As such, had the judges been presented with guidelines or an interface that explicitly asks them for judgments regarding effort, such incorrect inferences could be eliminated. Thus, the results presented here must be taken as a lower bound on the level of improvement one could obtain by incorporating effort together with relevance when judgments are obtained.

6. CONCLUSIONS

The concept of relevance sits at the core of information retrieval. Hence, much research has been devoted to defining and understanding relevance. There has been some previous work that conceptually distinguishes relevance from utility of a document to an actual user and argues that an evaluation metric should be utilitarian in nature [32]. However, relevance has been the main focus of most information retrieval researchers, implicitly assuming that utility and relevance are almost the same concepts.

In this paper, we propose a user model that shows how user behavior and expectation could be different than that of judges when reading a document. Based on this user model, we argue that the utility of a document to an actual user is highly affected by the effort the users need to find and consume relevant information, which is currently ignored by the relevance assessors. We further validate our hypothesis by showing that features related to readability of a document play an important role in the possible mismatch between relevance of a document versus its utility to an end user. We also propose a mechanism that uses the relevance, dwell time and judging time information on a document to infer the utility of the document to an actual user. We show that the utility judgments obtained in this way are much closer to the expectations according to our user model.

Overall, our results suggest that *effort* plays an important role in user satisfaction and that effort should be considered together with relevance when the quality of retrieval systems are evaluated. Our results also show that features related to

the effort to find/consume relevant information (e.g., readability level of the document, etc.) could be used as ranking features when retrieval systems are designed.

Given this, we argue that the guidelines given to relevance assessors should be changed and the assessors should provide judgments both in terms of the effort required to find relevant information in a document, as well as the relevance of the document. In a live retrieval system for which we have document-level data, this could be done by showing the median dwell time information to the judges. By understanding different search tasks and also the cases in which some users will quickly back out from a page, it could be possible to set judging guidelines that help judges identify high effort versus low effort documents. In the future, we plan to devise new judging mechanisms through which it is possible to get judgments on effort as well as relevance so that retrieval evaluation gets closer to evaluating user satisfaction.

7. REFERENCES

[1] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: Does the test collection predict users' effectiveness? In *Proc. SIGIR*, Singapore, 2008.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. SIGIR*, Salvador, Brazil, 2005.

[3] L. Azzopardi. The economics in interactive information retrieval. In *Proc. SIGIR*, pages 15–24, New York, NY, USA, 2011. ACM.

[4] L. Azzopardi. Modelling interaction with economic models of search. In *Proc. SIGIR*, pages 3–12, New York, NY, USA, 2014. ACM.

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

[6] J. C. Brown and M. Eskenazi. Student, text and curriculum modeling for reader-specific document retrieval.

[7] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proc. SIGIR*, pages 387–394, Singapore, 2008. ACM.

[8] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: A comparison to eye tracking. In *Proc. SIGIR*, Boston, USA, 2009.

[9] B. Carterette, P. D. Clough, E. Kanoulas, and M. Sanderson. Report on the ecir 2011 workshop on information retrieval over query sessions. *SIGIR Forum*, 45(2):76–80, Jan. 2012.

[10] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proc. CIKM*, Glasgow, UK, 2011.

[11] B. Carterette, E. Kanoulas, and E. Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proc. CIKM*, Maui, USA, 2012.

[12] B. Carterette and I. Soboroff. The effect of assessor error on ir system evaluation. In *Proc. SIGIR*, Geneva, Switzerland, 2010.

[13] P. Chandar, W. Webber, and B. Carterette. Document features predicting assessor disagreement. In *Proc. SIGIR*, Dublin, Ireland, 2013.

[14] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40. ACM, 2001.

[15] W. S. Cooper. On selecting a measure of retrieval effectiveness, part i. *JASIST*, 24(2):87–100, 1973.

[16] A. P. De Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. *RIAO Conference Proceedings*, pages 463–473, 2004.

[17] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proc. WSDM*, New York, USA, 2010.

[18] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proc. SIGIR*, Singapore, 2008.

[19] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005.

[20] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005.

[21] Q. Guo and E. Agichtein. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *Proc. WWW*, Lyon, France, 2012.

[22] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *Proc. SIGIR*, Athens, Greece, 2000.

[23] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2), 2007.

[24] G. Kazai, N. Craswell, E. Yilmaz, and S. Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proc. CIKM*, Maui, USA, 2012.

[25] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proc. SIGIR*, Sheffield, UK, 2004.

[26] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. In *ACM SIGIR Forum*, volume 37:2, pages 18–28. ACM, 2003.

[27] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, Sept. 2003.

[28] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM*, 2014.

[29] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[30] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proc. SIGIR*, Dublin, Ireland, 1994.

[31] F. Radlinski and T. Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Proc. AAAI*, 2006.

[32] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. *JASIST*, pages 321–343, 1975.

[33] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *JASIST*, 58(13):1915–1933, 2007.

[34] F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proc. SIGIR*, Beijing, China, 2011.

[35] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proc. SIGKDD*, Paris, France, 2009.

[36] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, Portland, USA, 2012.

[37] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results? In *Proc. SIGIR*, New Orleans, USA, 2001.

[38] R. Villa and M. Halvey. Is relevance hard work?: Evaluating the effort of making relevant assessments. In *Proc. SIGIR*, pages 765–768, 2013.

[39] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.

[40] K. Wang, T. Walker, and Z. Zheng. Pskip: Estimating relevance ranking quality from web search clickthrough data. In *Proc. SIGKDD*, pages 1355–1364, Paris, France, 2009.

[41] W. Webber. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop*, page 2, 2011.

[42] W. Webber, B. Toth, and M. Desamito. Effect of written instructions on assessor agreement. In *Proc. SIGIR*, Portland, USA, 2012.

[43] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proc. CIKM*, Arlington, USA, 2006.