Contextual and Dimensional Relevance Judgments for Reusable SERP-Level Evaluation

Peter B. Golbus
Northeastern University
Boston, MA
pgolbus@ccs.neu.edu

Imed Zitouni, Jin Young Kim,
Ahmed Hassan, Fernando Diaz
Microsoft
Redmond, WA
{izitouni,jink,hassanam,fdiaz}@microsoft.com

ABSTRACT

Document-level relevance judgments are a major component in the calculation of effectiveness metrics. Collecting high-quality judgments is therefore a critical step in information retrieval evaluation. However, the nature of, and the assumptions underlying, relevance judgment collection have not received much attention. In particular, relevance judgments are typically collected for each document in isolation, although users read each document in the context of other documents. In this work, we aim to investigate the nature of relevance judgment collection. We collect relevance labels in both isolated and conditional settings, and ask for judgments in various dimensions of relevance, as well as overall relevance. Then we compare the relevance metrics based on various types of judgments with other metrics of quality such as User Preference. Our analyses illuminate how these settings for judgment collection affect the quality and the characteristics of the judgments. We also find that the metrics based on conditional judgments show higher correlation with user preference than isolated judgments.

Categories and Subject Descriptors

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

Keywords

Evaluation, Contextual Relevance, Reusability

1. INTRODUCTION

In order to improve search engines, it is necessary to accurately measure their current performance. In recent years, much of the work on information retrieval evaluation has focused on user models [13][16][22] and diversity measures [1][15][24], which attempt to accurately reflect the experience of the user of a modern internet search engine. However, these new evaluation measures are only one aspect of the overall test collection evaluation methodology. That methodology relies on human relevance judgments for specific documents as input to these measures. While diversity measures and user models attempt to account for the interaction between documents, these human judgments are still collected under the assumption that the relevance of each document is independent of the relevance of the other documents. This simplifying assumption increases the reusability of test collections. However, recent work [10][27] has shown that the labels produced by assessors are dependent on other documents that they have seen.

 WWW'14, April 7–11, 2014, Seoul, Korea. ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2568015

Another issue with the current test collection evaluation methodology is that it is not clear whether it corresponds to actual user preference [2][3][25]. For this reason, alternative methodologies collect relevance labels at the Search Engine Result Page (SERP)-level, rather than the document level [4][28]. However, these SERP-level judgments, while potentially more accurate, are not **reusable**—they cannot be used to provide any information about future rankers.

The goal of this work is to explore the potential for incorporating contextual, **conditional relevance** into the existing test collection construction methodology without sacrificing reusability—its key advantage over online evaluation. While the diversity framework is an attempt to solve this problem, it suffers from several drawbacks. One is that judgments are made with respect to subtopics, defined in advance, that are often quite artificial [18]. Another is that all redundancies are treated equally. For example, a document that provides a general overview of a subject that is then followed by a document that provides additional detail on a specific aspect of the subject may be of more utility than either document by itself. However, the diversity framework will penalize this combination of documents. In our framework, we ask human assessors to tell us whether and to what extent document utility increases or decreases in specific contexts.

In this study, we focus on collecting the conditional relevance of documents in the context of a single, conditioning document: users are asked to read a document, and then bear it in mind when providing a relevance assessment for a second document. This gives us two sets of labels: traditional *isolation* labels, where users are asked to judge documents independent of any context, and *conditional* labels of the utility of a document in the context of the previous document. We expect the actual utility of a document to be reflected by some combination of these two labels. We show how to use these combined labels in traditional evaluation, as well as defining a new evaluation measure, **Contextual Cumulative Gain**, specifically created to make use of these conditional labels. We show that using a combination of isolation and conditional document labels increases the correlation with Search Engine Result Page (SERP)-level user preference.

The necessity of judging each document with respect to multiple context documents raises the number of required judgments, already the limiting factor in test collection construction, by an order of magnitude. To limit this growth, we propose the use of machine learning techniques to limit the increase in the required number of judgments to $O(n)$.

In addition, users were asked to label documents along a variety of relevance "dimensions" or "aspects" [21], such as topicality and freshness. This provides a better understanding of the impact

of conditioning on relevance: Section 5.1 shows how these aspects affect relevance assessments.

The main contributions of this work are as follows. (1) We conducted a user study in which we collected aspectual relevance labels for web documents in condition and in isolation from which we learned that: topicality is a necessary but not sufficient quality for a document to be labeled highly relevant, our aspects tended to have two "clusters"—one representing topicality and the other representing reliability, and that judges can be made more confident in their labels through the use of conditioning documents. We also demonstrate (2) a new evaluation methodology that collects conditional relevance labels with minimal additional overhead, and that incorporates those conditional labels into evaluation measures; as well as introducing a new measure: **contextual cumulative gain**, which is more highly correlated with user preference as indicated by SERP-level evaluation labels.

This paper proceeds as follows: after a brief discussion of related work (Section 2), we describe our evaluation methodology (Section 3). We provide details about our conditional relevance labels (Section 3.1), and then demonstrate how these labels will be predicted from context (Section 3.2). Finally, we show how these labels can be used in offline evaluation measures as well as defining a new evaluation measure designed to make use of these labels. Experimental setup, including the **user preference** methodology, is presented in Section 4. Finally, Section 5 shows obtained results, including the increased correlation between our methodology and user preference (Section 5.2), and the validity of our label prediction process (Section 5.3).

2. RELATED WORK

The concept of relevance has been extensively studied in IR. It is widely recognized for being multi-faceted and subjective. Human judgments are known to be influenced by various situational, cognitive, perceptual and motivational biases [7][23] as well as by document variables, judgment conditions and scales, and personal factors [26]. Novelty and diversity [2][11], which considers the interaction of multiple results, have gained increased attention with the launch of the diversity task at the TREC Web track [14]. This task required participating systems to retrieve a ranked list of documents that collectively satisfied multiple information needs, explicitly defined by the subtopics of a given test topic. However, while the interaction between documents is explicitly modeled, the retrieved documents are still assessed separately for each subtopic using a traditional judging procedure and binary relevance. The work which is most similar to our own is due to Chandar and Carterette [11][12]. The authors extended previous work using preference judgments rather than absolute judgments [9] to evaluate novel document retrieval methods. In their experiments, users gave preferences between documents given an already observed document. In this way, the authors were able to show that many assumptions underlying common diversity evaluation measures are false.

Recent work has shown that the relevance labels assigned to documents are influenced by documents previously seen in the same session [10][27]. A range of alternative IR evaluation methods have been proposed in recent years, aiming to go beyond the traditional methods that treat each retrieved document in isolation. For example, Bailey et al. [4] proposed a method that allows investigating aspects such as coherence, diversity and redundancy among the search results displayed in a SERP. Thomas and Hawking [28] proposed a preference method that displays two sets of search results user preference and asks users to indicate which side they preferred. In their experiments, comparing Google's first and second page results, they reported high levels of accuracy in users preferring the top ranked results. These results were extended in [21], which investigated the underlying criteria and relevance dimensions upon which user preference decisions may rest. We follow this work in our aspectual studies designed to understand the impact of conditioning.

3. EVALUATION METHODOLOGY

The current evaluation methodology consists of two parts: 1) a test collection—itself consisting of a corpus of documents, a set of queries, and relevance judgments describing the relationship between a subset of the documents and the queries; and 2) one or more target evaluation measures. With these elements, one can assess the performance of an information retrieval system with regard to a variety of tasks, *e.g.* ad hoc search, knowledge extraction, etc., depending on the type of relevance labels and measures used. Our methodology is novel in three ways: (1) we use "conditional" relevance labels that are collected in context; (2) we predict the labels that are missing, since it is not possible to collect all required conditional labels; and (3) we propose evaluation metrics that incorporate these collected and predicted conditional relevance labels. The collection and prediction of relevance labels are discussed in Section 3.1. Section 3.2 addresses the training data and machine learning techniques we use to predict missing labels, and Section 3.3 describes the use of these labels for evaluation, and introduces our new measure, Contextual Cumulative Gain.

3.1 Conditional Relevance Labels

Imagine reading a highly informative document that provides a high quality overview of the topic the user is searching for. This is clearly a document with high utility for the user. Now imagine the next document a user is presented with is another high quality overview that mostly covers the same information. While, in isolation, both of these documents would receive high relevance scores, in context, the second document has very little conditional utility, and should therefore receive a smaller conditional relevance grade. In contrast, a highly specific document that expands upon a single aspect of the first document may conceivably have an increased contextual utility. While the first scenario may be dealt with through the use of subtopic labels and diversity evaluation measures, the second example would also be penalized in that framework. Further, the diversity framework treats all redundancies equally, whereas some documents may interact with each other more strongly. Rather than try to predetermine a user model that applies equally to all contexts, we solicit this information from a judge directly.

We assume that users interact with each document in ranked list order and that each document is experienced in the context of all preceding documents. Since this would require a factorial number of judgments, we simplify this model and focus on document pairs: we ask users to provide *isolation* relevance labels for individual documents with no context, $g(d_i)$, we also ask users to provide us with *conditional* labels for each document in the context of the previous document in a hypothetical ranked list, $g(d_i|d_{i-1})$.

3.2 Predicting Relevance Grades

The number of adjacent pairs of documents within all hypothetical ranked lists is still too large to collect. Therefore, we collect a limited number of judgments based on candidate lists and use them as training data to predict the remaining labels necessary for

evaluation. We begin with ranked lists produced by actual search engines. For each query, we collect isolation judgments for each of the top five documents, as well as conditioning the documents at ranks two through five upon the document at rank one. These labels serve both as training data for our prediction process, as well as a baseline to compare against. We also want to observe the impact of conditioning on documents of different quality ranges to make training data more representative. Therefore, we randomly assign each query to one of two sets for additional judgments. For roughly 3/5ths of the queries, we evaluated the first four documents on the fifth document. For the remaining 2/5ths we collect isolation judgments for the documents at ranks six through ten, as well as labels for the documents at ranks seven through ten conditioned upon document at rank six.

We use a multinomial logistic regression classifier to predict the relevance grades of unlabeled pairs. Since training data is sparse, we augment it using intuition from language modeling [20]. If we consider our conditional relevance grades as equivalent to bigrams, or word pairs, then we use a process inspired by smoothing, which is used to account for the fact that a corpus will not contain all valid word pairs. Since we have relevance grades for every document in isolation, we compute the utility of each document as the linear interpolation between the isolation grade and the predicted context grade:

$$(1-\alpha)g(d_i) + \alpha g(d_i|d_{i-1}) \quad (1)$$

where α controls the weight given to the conditional labels, ranging from 0 (for pure isolation labels) to 1 (pure contextual labels). We interpret $g(d_1|d_0)$ as being equivalent to $g(d_1)$. If $g(d_i|d_{i-1})$ was collected, then we use it in Formula 1. Otherwise, we use a predicted value.

3.3 Evaluation Metrics

In this section, we first describe existing baseline models (Section 3.3.1) before introducing our contextual evaluation models (Section 3.3.2).

3.3.1 *Baseline Evaluation Models*

Evaluation models can be understood in terms of the hypothetical user that they describe. For example, *normalized discounted cumulative gain* (nDCG) [19], describes the experience of a user who browses to a pre-determined rank k, deriving utility from each document in an amount proportional to the document's relevance grade and inversely proportional to the rank at which the document is encountered. We first define *discounted cumulative gain* (DCG).

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{2^{g(d_i)}-1}{\log(i+1)} \quad (2)$$

Since the range of DCG will vary from topic to topic, we normalize these scores so that an average can be computed. Normalization is performed with regard to the maximum possible DCG of an ideal ranked list.

Craswell et al. [16] introduced the Cascade model of user behavior. In this model, a user is assumed to browse documents in order until reaching a satisfying document. This implies that if a user reaches rank k, then none of the $k-1$ documents ranked before it were satisfying. Therefore the probability of a user reaching rank k depends on the relevance grade of that document, and each of the previous documents in the list.

Chapelle et al. [13] developed an evaluation measure, *expected reciprocal rank* (ERR), based on the Cascade model. Let R_i denote the probability that a user will find the document at rank i to be satisfying. One common function used to estimate this is

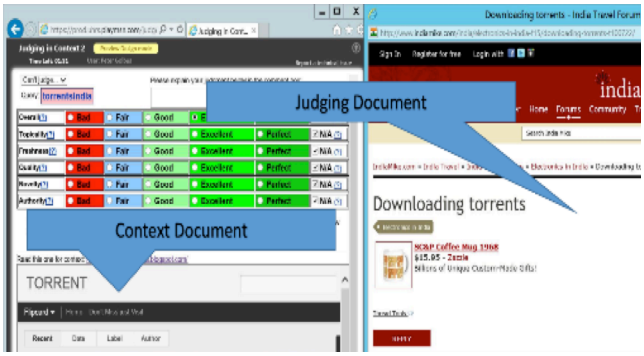

Figure 1: Judging interface used to collect document labels in condition. Context document is left blank for isolation.

$$R_i \approx \frac{2^{g(d_i)} - 1}{2^{gmax}} \quad (3)$$

where $gmax$ is the maximum possible relevance grade in the collection. Then in the Cascade model, the probability that a user will terminate his or her search at rank k is

$$P(k) = R_k \prod_{i=1}^{k-1} 1 - R_i \quad (4)$$

and the expected reciprocal rank at which a user will terminate his or her search is

$$\text{ERR} = \sum_{k=1}^{\infty} \frac{R_k}{k} \prod_{i=1}^{k-1} 1 - R_i. \quad (5)$$

3.3.2 *Contextual Evaluation Models*

These measures described earlier can easily be made to use our conditional labels by simply replacing $g(d_i)$ with our interpolated prediction model (Equation 1). In addition, we define a new measure *contextual cumulative gain* (CCG), specifically designed with these conditional labels in mind, which we show to be more correlated with user preference in Section 5.2.

In our model, we interpret the logarithmic discount function as a stopping "probability" in a manner similar to that of Carterette [8]. Rather than interpreting k as the pre-determined stopping rank, we view the logarithmic rank discount to be proportional to the likelihood that the user stops at each rank. In our model, if a user stops at rank 3, for example, then, using our conditional labels, the user will derive utility from the first document, from the second document conditioned on the first document, and from the third document conditioned on the second. This will occur with some probability proportional to $\frac{1}{\log(3+1)} = \frac{1}{2}$. We define CCG@k as follows

$$\text{CCG@k} = \sum_{i=1}^{k} \frac{\sum_{j=1}^{i}(1-\alpha)g(d_j)+\alpha g(d_j|d_{j-1})}{\log(i+1)}. \quad (6)$$

This can also be expressed in a single summation as

$$\text{CCG@k} = \sum_{i=1}^{k}(k-i+1)\frac{(1-\alpha)g(d_i)+\alpha g(d_i|d_{i-1})}{\log(i+1)}. \quad (7)$$

We demonstrate Equation 6 as it makes the intent behind the user model clearer. Since all users will interact with the document at rank one while only relatively few users will interact with the document at rank k, the document at rank one contributes much more strongly to CCG than it does to the standard DCG metric. There is also the hypothetical advantage that since better

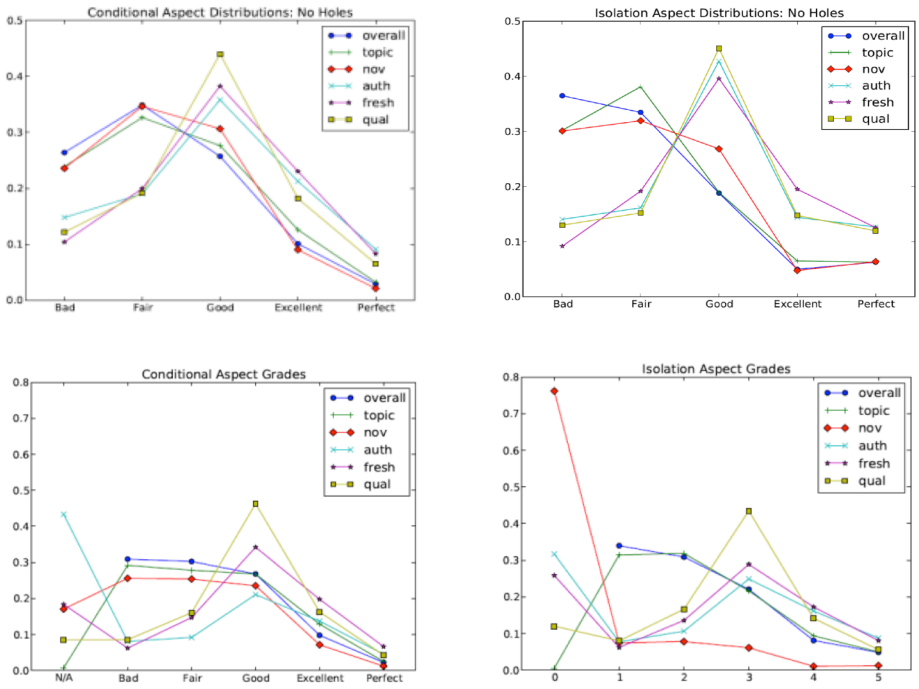

Figure 2: Multinomial distribution over aspect grades in condition (left) and isolation (right). Distributions above are over documents that have non-N/A labels for all aspects, below are over documents with overall grades that are not N/A.

documents will appear before worse documents, later documents in the list should have little conditional utility.

4. EXPERIMENTAL SETUP

We present here the user study setup (Section 4.1), and how we compared our methodology to user preferences (Section 4.2).

4.1 Judging Interface & Guideline

The users in our study were professional judges that had extensive training in producing standard relevance judgments of web documents. Our task differed from the standard task in two ways: (1) rather than providing us with a single overall label, judges were asked to provide multiple labels, providing us with a deeper understanding of the impact of our methodology; and (2) judges were asked to provide these grades for documents in isolation and in condition.

Judges were asked to assign grades (bad, fair, good, excellent, perfect) to documents with respect to each of the following: overall, topicality, novelty, authority, freshness, and quality. Judges were also allowed to mark an aspect as Not-Applicable (N/A), indicating that either the judge was unable to ascertain the correct label for the document, or was unable to apply the aspect to the query.

Judges were presented with a single user interface (Figure 1) when asked to judge documents in isolation and in condition. With the exception of additional judgments collected for the purposes of inter-assessor disagreement experiments, each query was judged by a single assessor, whether in condition or in isolation. That way, there are no inter-assessor effects between conditional and isolation judgments. The judgment in isolation and its counterpart in context are conducted in different period of time to reduce bias. Aspects were defined for the judges in the following way:

Overall: Given that the user just read the first web page, in general, how satisfied would the user be if they read the new web page in context?

Topicality (topic): Given that the user just read the first web page, does the new document seem useful to a likely intent for this query in context?

Novelty (nov): This aspect asks about the information covered in the two web pages. Given that the user just read the first web page, does the new web page provide entirely new information with no overlap, or is it completely redundant in context?

Freshness (fresh): Given that the user just read the first web page, does the new web page seem to contain the most up-to-date information about the topic in context?

Authority (auth): Given that the user just read the first web page, does the new web page appear to have a level of credibility appropriate to the topic in context?

Quality (qual): Given that the user just read the first web page, does the new web page seem well-written and well-organized in context?

Although these definitions refer to context only, judges were given no additional instructions about the meaning of these aspects in isolation. This is especially interesting with regards to novelty, as it is not clear how a judge should interpret this without having a reference document. We collected relevance labels over 457 queries—sampled from the logs of a major commercial search engine—for which we had user preferences between two SERPs. Of these queries, 270 queries had labels for the first five documents in isolation, and were conditioned on at least the documents at ranks one and five. We also had 188 queries[1] with

[1] One query was accidentally judged twice.

labels for the first ten documents in isolation, the first five conditioned on the document at rank one and the second five conditioned on the document at rank six.

Several queries were judged by multiple judges for use in inter-assessor experiments. We found that the agreement between judges was similar in isolation and in condition. On average, judges had a Cohen's Kappa agreement of 0.412 in isolation and 0.367 in condition, and a Jaccard coefficient of 0.568 in isolation and 0.548 in condition.

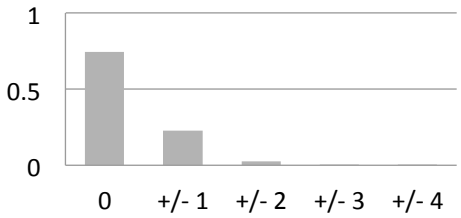

Figure 3: Probability of grade change from isolation to condition by magnitude.

4.2 Comparison with User Preference

Our goal in this work is to compare out methodology against user experience. One of our primary proxies for user experience is what we refer to as **user preference** [21][28]. User preference is measured by asking users to make side-by-side comparisons of rankers. Users are shown two SERPs and asked if one is preferable to the other, and by how much. These labels are mapped onto directional preferences {-1,0,1} in one of two ways: **weak user preference**, in which labels are divided such that any preference, including a slight one, is sufficient; and **strong user preference**, in which labels are divided such that slight preferences are considered to be ties.

For each query, the output of two commercial search engines were randomly assigned as either the "left" system or "right" system. Preferences between the "left" and "right" systems were collected from 5 different assessors. The label we use is the modal directional preference. In the data that we collected, user's had a weak preference on roughly 91% of queries, but a strong preference on only 23%.

5. RESULTS

Section 5.1 details our observations with regard to dimensional relevance, while Section 5.2 demonstrates the correlation between conditional labels and user preference. Section 5.3 describes the results of our label prediction process.

5.1 Dimensional Relevance

The goal of this aspect of the study is to understand the behavior of our assessors in terms of the impact of various relevance "dimensions," and its impact on conditioning when compared to isolation. The following points are discussed in detail below. Our findings indicate that users are more likely to have definitive opinions in condition than isolation (**aspect coverage**). Also, users assign different grades in isolation than in condition, yet in the same overall proportion (**aspect distribution**). Next we discuss observed patterns of document context affecting relevance (**impact of conditioning**). Finally, we show that topicality and authority represent two separate yet important factors in determining overall document quality by examining the correlation between aspects (**aspect correlation**), as well as the extent to which the various aspects can be used to predict overall grades (**predictive power**). In these discussions, we exclude documents with overall labels of "N/A." These documents are primarily spam, and their impact on assessors is uninteresting. However, we do consider these documents in our discussion of predictive power, as we wish to leverage as much information as possible.

Aspect Coverage: Our first finding was that users were more likely to provide aspect labels other than "N/A" in context than in isolation. In isolation, only 18.4% of documents had non-"N/A" labels for all aspects, while 42.6% had non-"N/A" labels for all aspects in condition. A large part of the low number in isolation is due to the fact that the novelty aspect is hard for judges to assess when document is shown in isolation. However, Table 1 shows that, with the exception of authority, judges were more likely to provide non-"N/A" labels for each aspect. The table excludes documents that had "N/A" overall labels.

Table 1: Percentage of documents covered for each aspect in isolation and in condition.

	Isolation	Condition
Topicality	99.4%	99.3%
Novelty	22.7%	81.8%
Authority	68.5%	55.0%
Freshness	74.9%	80.4%
Quality	88.7%	90.9%

From this we conclude that making judgments in condition can help judges that are uncertain reach specific decisions about the quality of documents.

Aspect Distribution: We also discovered that judges tended to use labels in the same general proportions in isolation as they did in condition. Figure 2 shows multinomial distributions for each aspect in isolation and in condition. Since each aspect has a different set of documents with "N/A" labels, the distributions above are restricted to documents with "no holes," which is to say that they have non-"N/A" labels for all aspects. The distributions below show all documents with non-"N/A" overall labels. We observe that aspects tend to fall into two "clusters" with highly similar distributions: (1) a cluster with overall, topicality, and novelty; and (2) a cluster with authority, freshness, and quality.

Given that these distributions are so similar in isolation and in condition, it is reasonable to wonder whether document grades were impacted by conditioning. Figure 3 shows that probability of various grade changes by magnitude between isolation and conditional grades. For example, if a document had an isolation and conditional grade of "good," it would be represented in the 0 column. If it had a conditional grade of "fair" and an isolation grade of "good," or vice versa, it would contribute to the +/-1 column. Our results show that while the majority of the time ($\approx 75\%$) there is no change in document grade, a sizable number of documents ($\approx 22\%$) change by one grade, and some ($\approx 3\%$) change by 2 or more. This indicates that there are some documents that a user may find either fair or excellent depending on the context. In future work, we tend to explore the magnitude of this change relative to expected intra-assessor disagreement, which is beyond the scope of this paper.

Impact of Conditioning: In general, we did find reasonable patterns of changes in overall grade labels. For example, the use of conditioning documents was able to provide judges with a greater understanding of query results, similar to priming [27]. There are contexts that can make it clear that a result is relevant in

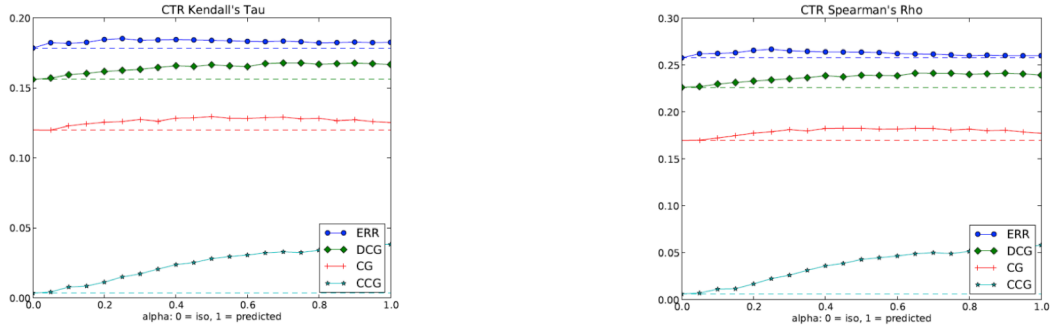

Figure 4: Rank correlation between queries sorted by mean click-through rate and offline evaluation measures. The term α controls the weight of conditional labels.

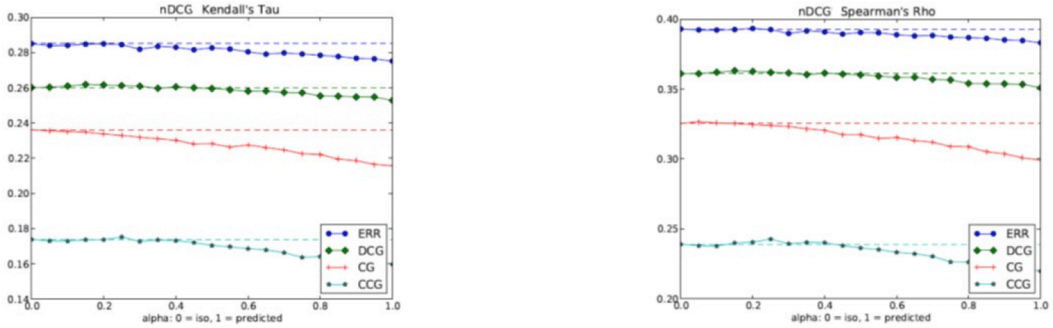

Figure 5: Rank correlation between queries sorted by nDCG using labels generated from click data and offline evaluation measures. α controls the weight of conditional labels.

a way that may not be clear to a judge in isolation. Consider the query "rick james dave chappele." A judge may not be aware of the comedy routine, and may therefore not recognize images of Dave Chappele dressed as Rick James as relevant. However, conditioning on a video of the routine makes this apparent.

Another example of the impact of conditioning is the scope of the results. Documents may appear very specifically targeted towards the user intent in isolation, yet a conditioning document can reveal that they are still not specific enough. Consider the query "Lionel engines o 27." Lionel is a producer of toy train engines, and 0-27 is a standard toy train engine size, or "gauge." The user is searching for a specific brand of model train engine in a particular size. A listing of Lionel train engines for sale may appear to be highly targeted. However, if this listing is conditioned on model train parts that are specifically O-27 gauge, the judge (or user) may find their opinion of the utility of the listing to be reduced.

A final example is about document quality. Imagine that the user is searching for the answer to a specific question. If the answer is contained within a document that is poorly written and has a lot of irrelevant text, this document may still get a relatively high grade. However, this grade will be far lower if conditioned on a well-written document that is more on topic.

Aspect Correlation: We next turn our attention to the correlation between the various aspects. Did asking the users for additional relevance dimensions provide additional information, or was the same label applied to all aspects? One way to answer this is to compare the document rankings induced by the various aspect grades against the ranking induced by the overall grades. As each aspect has a different set of documents with non-N/A labels, we must ensure that we rank a consistent set of documents. In Table 2**,** we rank the subset of documents that have non-N/A labels for *all* aspects and report the Spearman's Rho rank correlation. Of all the aspects, topicality is clearly the most correlated with overall grades. This shows that topicality is necessary but not sufficient for a document to be highly relevant. Also, we note that it is more highly correlated in isolation than in condition. We take this as further evidence that conditioning documents do have an impact on overall grades. Next, we explore whether various aspects change in the same way. For example, it may not be the case that documents tend to have similar authority and novelty grades, but that authority and novelty may increase under the same conditions. In Table 3, we rank documents by the magnitude difference between conditional and isolation grades for each aspect, i.e. $\boldsymbol{g(d_i|d_j) - g(d_i)}$**.** For each pair of aspects, we rank all documents that have non-N/A labels for that *pair* of aspects. Therefore, the documents under consideration between novelty and authority may not be the same as those being considered between freshness and quality. All documents have valid overall labels. This table supports our conclusion that overall, topicality and novelty grades form one "cluster" and authority, freshness, and quality grades form a separate "cluster."

Predictive Power: To further analyze the relationship between aspects and overall grades, we also consider their predictive power—if I only knew the aspect labels, how often would I be able to predict the correct overall label?

We use a multinomial logistic regression classifier on a random 90/10 train/test split of all of our isolation labels—not just those with non-N/A labels as in Table 3—to see which combinations of aspects can be used to most accurately predict the overall label. If we use all aspects, we achieve a (micro-)accuracy of 0.9494. If we consider only topicality and authority, this actually increases to 0.9570. Using only topicality provides an accuracy of 0.5494,[2]

[2] Note that there are six grades: N/A, Bad, Fair, Good, Excellent, and Perfect and hence a 0.5 accuracy is not equivalent to random chance.

while considering only authority provides a much higher accuracy of 0.8051. This high correlation between authority and overall performance in isolation is also confirmed by the work of Kim et al. [21].

The same experiment performed on our conditional labels provides consistent results. Using all aspects, we achieve an accuracy of 0.872, whereas using only topicality and authority yields a virtually identical accuracy of 0.866. If we use topicality as our only feature, we achieve an accuracy of 0.572. Using only authority yields an accuracy of 0.674. This is still higher than topicality, but not as much so as in isolation.

We believe that models trained on authority alone are more accurate then models trained on topicality because of the vast number of poor quality documents. Even though topicality is more correlated across all relevance grades, low and N/A authority grades are very highly correlated with low overall grades.

Table 2: Spearman's Rho correlation between aspects and the conditional / isolation grades. Docs have all aspects labeled.

Conditional Aspects	Conditional Overall	Isolation Aspects	Isolation Overall
Topicality	0.865	Topicality	0.923
Isolation Overall	0.828	Novelty	0.863
Novelty	0.69	Quality	0.484
Quality	0.511	Authority	0.48
Freshness	0.469	Freshness	0.444
Authority	0.41		

Table 3: Spearman's Rho correlation between the magnitude difference between conditional and isolation grades, e.g. $g(d_i|d_j) - g(d_i)$. Docs have labels for each pair of aspects.

	Overall	Topic	Nov	Qual	Fresh	Auth
Overall	—	0.607	0.378	0.190	0.189	0.254
Topic		—	0.373	0.172	0.165	0.215
Nov			—	0.231	0.149	0.243
Qual				—	0.280	0.307
Fresh					—	0.303
Auth						—

5.2 Correlation with User Preference

In this section, we explore how the use of conditional labels affects the correlation between offline evaluation measures and actual user experience. We have two proxies for user experience: (1) online evaluation measures and click behavior (Section 5.2.1) and (2) user preference as describe in Section 4.2 (Section 5.2.2).

5.2.1 Online Metrics

Our click data consists of a sample of hundreds of thousands of search sessions from the logs of a commercial search engine. We used a total of four weeks of interaction logs from March 2013. These log entries include a unique user identifier, and a timestamp for each page view. They also include all queries and clicked Web pages. Intranet, secure (https) URL visits, and any personally identifiable information were excluded from the logs. We tried to determine if the notion of query difficulty [4] captured by click measures was the same as those collected by online measures. For each query, we compute a score using each of our offline metrics and an online click-based measure. We then compare the ranking of queries induced by the offline measures and the ranking induced by the click measure. These rankings indicate those queries on which our ranker was most successful—if our offline measures are truly indicative of user satisfaction, then the queries with high evaluation scores should also be the queries on which users were most satisfied, as measured by click behavior.

To measure click behavior we use Mean Click-Through Rate and nDCG using labels derived from click data [29]. Mean Click-Through rate is the average ratio between the number of times documents were presented to users (in any position) and the number of satisfactory clicks[3] on those. Note that this is a measure on a *set* of documents, and will be the same for any *list* of those documents. To compute nDCG, we assign document labels based on clicks. A document has a label of two if it received at least one satisfactory click (at any rank), a label of one if it received at least one click of any kind, and a label of zero if it received no clicks whatsoever. All evaluation measures are computed at rank five.

Figure 4 shows the Kendall's Tau and Spearman's Rho list correlations between the queries sorted by mean CTR and the queries sorted by the various offline measures as alpha varies. We see that the correlation is small for all measures. We also see that there is a mild increase in correlation as α, the weight given to our contextual labels, increases. This is especially pronounced for our new evaluation measure, which is especially designed to make use of these labels. Figure 5 shows the list correlations between the query rankings of offline measures and user satisfaction as measured by nDCG using click based labels. As before, there is little change as α is varied, but in this case increasing the weight given to contextual judgments seems to decrease the correlation with user behavior.

Obtained results support our hypothesis that conditional labels, even on only a single document, are more indicative of user experience. The click data was collected from actual users interacting with many ranked lists, not just those used to generate the labels, i.e. we are comparing $g(d_i|d_j)$ against click data for document i, but document j did not necessarily appear before document i in any of the lists used to collect click data. If conditional labels were unimportant, than this effect should not matter—$g(d_i|d_j)$ should be equivalent to both $g(d_i)$ and the click information for document i no matter what lists the observations came from. However the correlation with CTR, a set based measure, increased while the correlation with the online variant of nDCG, a list based measure, decreased. This means that if the evaluation score of a list increased, then since it increases for at least one permutation of documents, it is likely to increase for a set measure. However, the offline evaluation score is independent of the online list-based measure—just because it increased in one ranked list doesn't mean anything about the behavior of the lists actually presented to users.

5.2.2 User preference

We measure the impact of conditional labels on evaluation as measured by User Preference (Section 4.2): given two rankings, does the preferred ranker also have a higher evaluation score? This comparison requires us to define a tie for each measure. Therefore, there are two parameters: α, the weight given to conditional labels, and the score threshold used to define a "tie."

[3]As a proxy for satisfaction, we define a "satisfied" click as a click with a dwell time of at least 30 seconds [17].

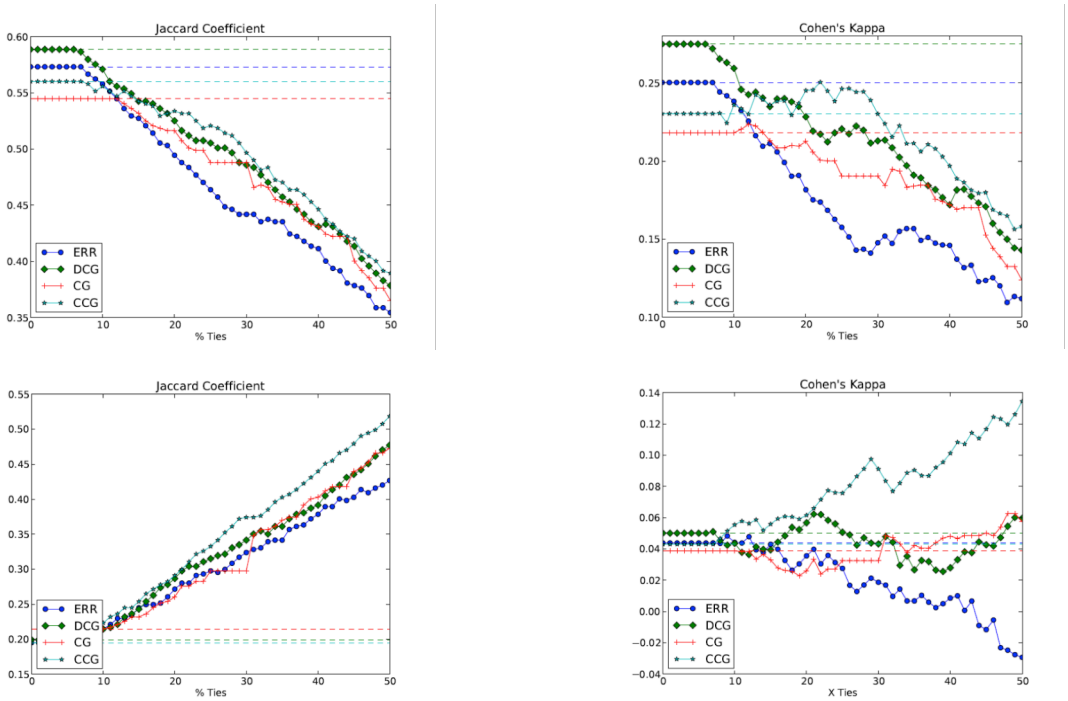

Figure 6: Agreement correlation between weak (upper plots) and strong (lower plots) user preference and offline evaluation measures as the threshold for ties is varied. α is fixed at 0.5.

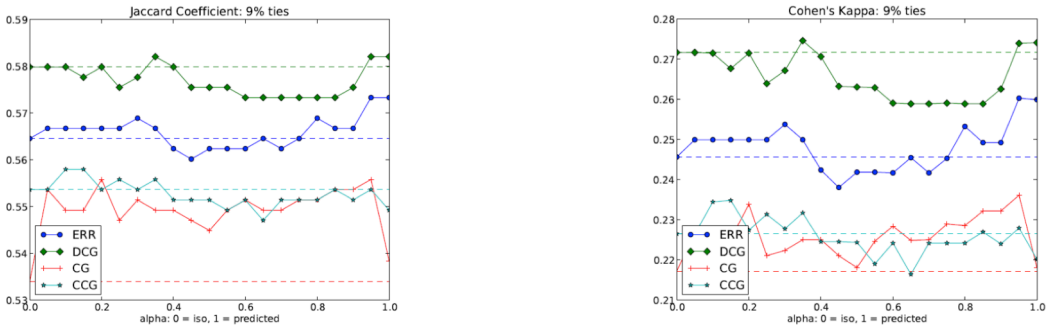

Figure 7: Agreement correlation between weak user preference and offline evaluation measures as α is varied. The tie threshold is such that the offline measures report approximately the same number of ties as user preference.

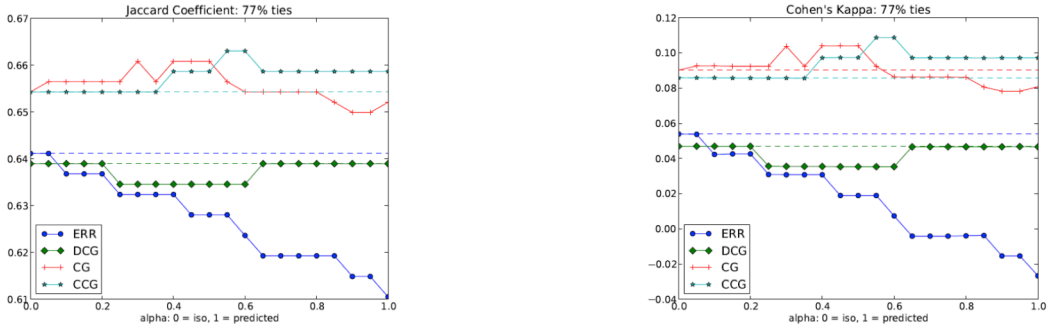

Figure 8: Agreement correlation between strong user preference and offline evaluation measures as α is varied. The tie threshold is such that the offline measures report approximately the same number of ties as user preference.

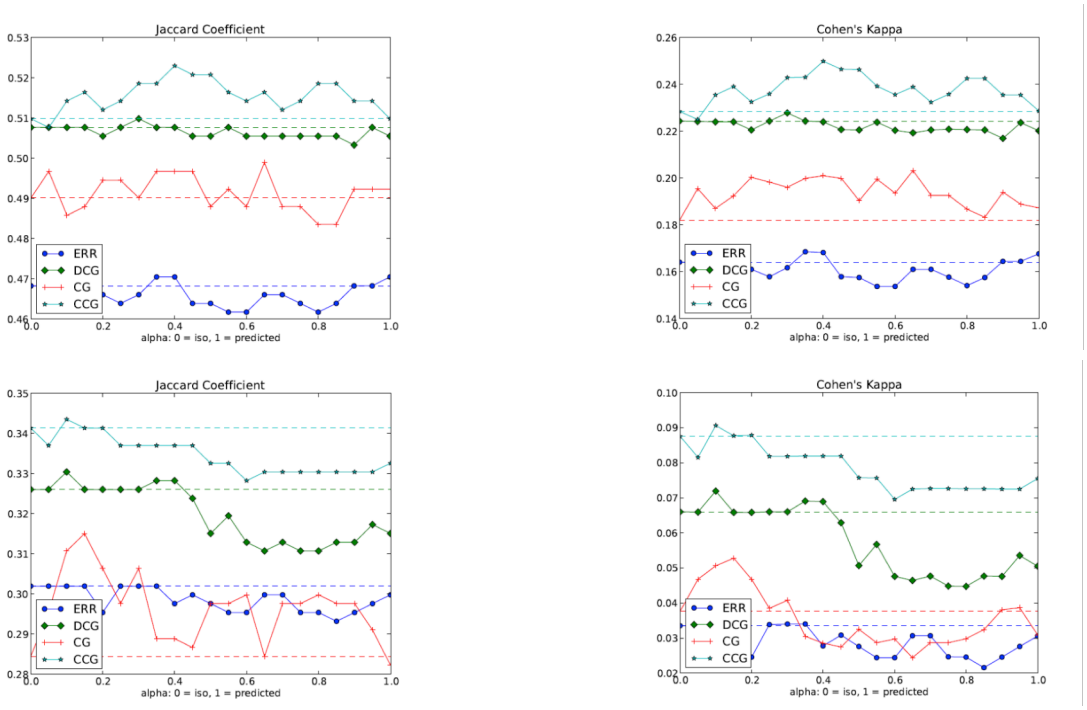

Figure 9: Agreement correlation between weak (upper plots) and strong (lower plots) user preference and offline evaluation measures as α, the weight given to conditional labels, is varied. Tie threshold is chosen so as to maximize simultaneous agreement with weak and strong user preference.

Thresholding: We are interested in the effects of α, which contains information about the utility of our conditional labels. However, we must choose tie thresholds in order to compare against User Preference. We choose a threshold based on the number of ties reported. If $x\%$ of queries have a difference of t or less, then choosing t as threshold will cause measure to report ties on $x\%$ of queries. With $\alpha = 0.5$, Figure 6 shows the Jaccard Coefficient and Cohen's Kappa between the measures and weak and strong user preference, respectively, as the percentage of queries that will be considered to be tied is varied. The dashed lines indicate the agreement with the smallest possible threshold (% ties equal to zero). Further, as each query is considered independently and not averaged, there is no need to normalize across queries; we report on the behavior of DCG, which is equivalent to nDCG. All evaluation measures are computed at rank five.

We note that plots in Figure 6 are initially horizontal. This is because a threshold of zero will force each measure to report some number of ties, *i.e.* there exists some minimum percentage of ties for each measure. The plots show us that the agreement varies greatly as we change the tie threshold. Further, the impact of changing the threshold is opposite between weak and strong user preference. Since more ties are reported by strong user preference, which treats slight preferences as ties, increasing the tie threshold increases the agreement with strong user preference and weakens the agreement with weak user preference. Therefore, we report our results under three different conditions:

1. **Weak User Preference**: First, we compare our methodology to user preference under conditions dictated by weak user preference, i.e. since users reported ties on 9% of queries, we set the threshold for each measure to report ties on ~ 9% of queries.
2. **Strong User Preference**: However, it is difficult for evaluation measures to be as sensitive as human assessors. To present our evaluation measures with a more reasonable task, we also use conditions dictated by strong user preference.
3. **Simultaneous Agreement**: Unfortunately, strong user preference reports a tie 77% of the time, meaning that simply reporting that any pair of rankers is tied on any query is likely to outperform any principled approach. Therefore, we also present results with the percentage of ties set to ~25%, maximizing simultaneous agreement with both weak and strong user preference.

Weak User Preference: Figure 7 shows the agreement between weak user preference and the measures with corresponding tie thresholds, with the dashed lines indicating the agreement when $\alpha = 0$, *i.e.* when we use pure traditional isolation labels. We observe that DCG has the highest agreement, and that the agreement goes down as we increase α, increasing the weight given to conditional labels. For all other measures, the agreement is maximized for some $\alpha \neq 0$, implying that there is some use of conditional labels that increases agreement. This increase is largest for cumulative gain, supporting our notion that using conditional labels provides something akin to a discounting factor.

Strong User Preference: Figure 8 shows the agreement with strong user preference. We observe that CG and CCG have the highest agreement, and that their agreement is maximized by giving roughly equal weight to isolation and conditional labels. However, notice that while the Jaccard Coefficient is reasonably large, denoting high agreement, the Cohen's Kappa is almost zero. This implies that given the preponderance of ties reported, none of

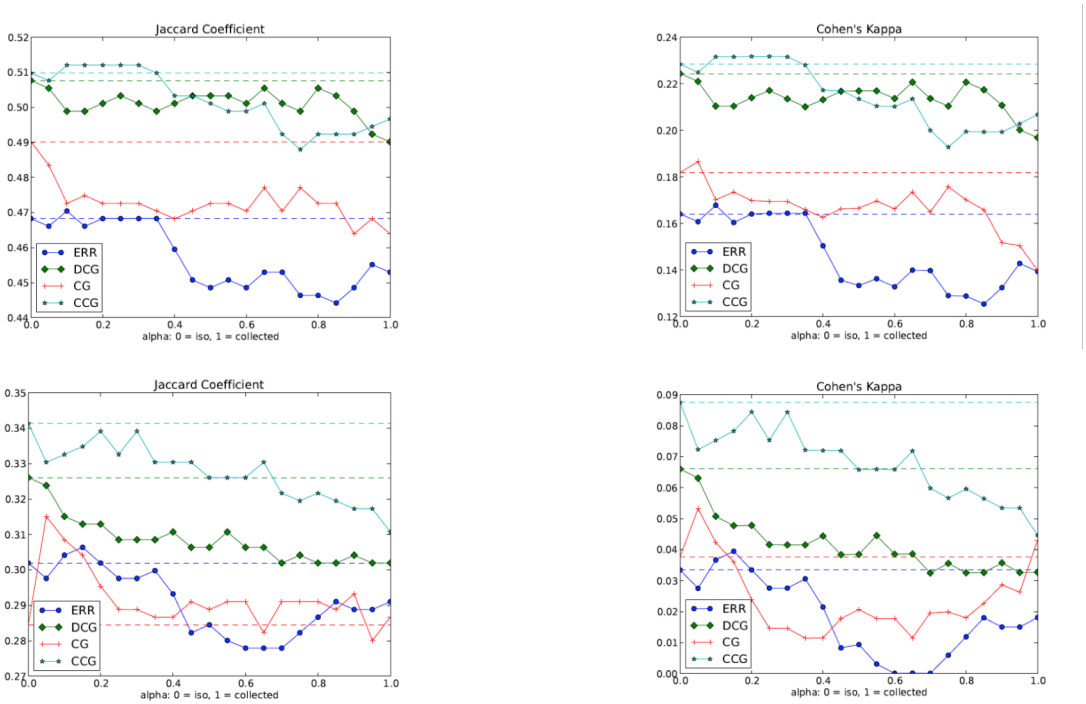

Figure 10: Agreement correlation between weak (upper plots) and strong (lower plots) user preference and offline evaluation measures using collected labels rather than predicted labels.

our measures are outperforming a baseline of simply reporting a tie on all queries, which would have a large agreement by chance.

Simultaneous Agreement: In this case, we maximize simultaneous agreement, which is a fairer test of the impact of our conditional labels. Figure 9 shows the agreement on weak and strong user preference when we set the threshold such that measures report ~25% of queries as being tied. In the case of weak user preference, CCG has the highest agreement, and giving some weight to conditional labels clearly improves the performance, as it also does for CG. DCG also has high agreement with weak user preference while ERR has the lowest. Both seem to be relatively unaffected by the use of conditional labels. When comparing to strong user preference (Figure 9, lower plots), CCG again has the highest agreement. The use of conditional labels seems to negatively impact all measures except for CG. However, as each measure can report ties only 25% of the time whereas the "correct" answer would be to report a tie on 77% of queries, there is a limit on how high the agreement can be.

5.3 Predicting Labels

As discussed in Section 3.2, we still required many more labels to perform evaluation at rank five than we actually collected. We used a multinomial logistic regression classifier using only the overall isolation labels $g(d_i)$ and $g(d_{i-1})$ as features to predict the label $g(d_i|d_{i-1})$. Our model achieved results of about 75% accuracy on a 90/10 train/test random split.

To validate our model, we compare it against evaluation using only labels that were collected directly. If we do not have a label for $g(d_i|d_{i-1})$, another option other than predicting it would be to use a similar label that we have collected. Here we use $g(d_i|d_1)$, which we collected for all documents, rather than a predicted value. This is similar to backoff smoothing [6] in language modeling, where if we have not seen a given bigram, we will check to see if our corpus contains the current word appearing after any recently encountered words.

Figure 10 shows the Jaccard coefficient and Cohen's Kappa with the threshold set to produce ties on 25% of queries. These results show that increasing α, the weight given to the context labels (now collected rather than predicted), unambiguously decreases the agreement between measures and user preference. This indicates that our machine learning approach does provide valuable information about the conditional utility of documents, and therefore can provide conditional labels beyond those collected by judges. As we intend to validate in future work, this implies that our approach can be used in practice without sacrificing reusability.

6. CONCLUSION

We conducted a user study in which we collected aspectual relevance labels for web documents in condition and isolation. This provided valuable insight into how user experience is effected by SERP-level context. Furthermore, we demonstrated that assessors can be made more confident in their labels through the use of conditioning. We also demonstrated a new evaluation methodology that incorporates conditional relevance labels with minimal additional overhead. Our new evaluation paradigm is more highly correlated with user preference as indicated by SERP-level evaluation labels. The results were consistent across parameter values when we compared against the strong preference judgment.

While user preference is a very informative measure, it is not **reusable**—if we wish to know if a user prefers a new system over an existing system, we cannot use any existing labels. This is, potentially, the true power of our methodology: since we predict new labels from the labels we have collected, in principle, we should be able to evaluate new systems without requiring new judgments to be collected.

7. REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. WSDM '09.

[2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough, The relationship between IR effectiveness measures and user satisfaction. SIGIR '07.

[3] Al-Maskari, A., Sanderson, M., Clough, P., & Airio, E. The good and the bad system: does the test collection predict users' effectiveness? SIGIR '08.

[4] Aslam, J. A., & Pavlu, V. (2007). Query hardness estimation using jensen-shannon divergence among multiple scoring functions. ECIR'07.

[5] Peter Bailey et al., "Evaluating search systems using result page context," in *IIiX*, 2010.

[6] Daniel Bikel and Imed Zitouni. Multilingual Natural Language Processing Applications: From Theory to Practice. Publisher Prentice Hall, Chapter 5, pages 169-198, 2012.

[7] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913-925, 2003.

[8] Ben Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. SIGIR '11.

[9] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais, "Here or There," in *ECIR*, 2008.

[10] Ben Carterette and Ian Soboroff. 2010. The effect of assessor error on IR system evaluation. SIGIR '10.

[11] Praveen Chandar and Ben Carterette. Using preference judgments for novel document retrieval. SIGIR '12.

[12] Chandar, P**.** and Carterette, B. Preference Based Evaluation Measures for Novelty and Diversity. *SIGIR'13*.

[13] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. CIKM '09.

[14] C. Clarke, N. Craswell, and I. Soboroff, "Overview of the TREC 2009 Web Track," in *TREC Proceedings*, 2009.

[15] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, Ian Mackinnon. Novelty and Diversity in Information Retrieval Evaluation. SIGIR '08.

[16] Nick Craswell, Onno Zoeter, Michael Taylor and Bill Ramsey. An experimental comparison of click position-bias models. WSDM '08.

[17] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.* 23, 2 (April 2005), 147-168.

[18] Peter B. Golbus , Javed A. Aslam , Charles L. Clarke, Increasing evaluation sensitivity to diversity, *Information Retrieval*, v.16 n.4, p.530-555, August 2013.

[19] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS,* 20(4):422-446, October 2002.

[20] Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recogniser. IEEE Transactions on Acoustics, Speech, and Signal Processing, 35(3), 400–401.

[21] Jinyoung Kim, Gabriella Kazai, and Imed Zitouni, Relevance Dimensions in Preference-based IR Evaluation. SIGIR '13.

[22] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems,* 27(1):2:1-2:27, December 2008.

[23] E. Pronin, "Perception and misperception of bias in human judgment," *Trends in cognitive sciences*, vol. 11, no. 1, pp. 37-43, 2007.

[24] Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. SIGIR '11.

[25] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas, Do user preferences and evaluation measures line up? SIGIR '10.

[26] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance," *JASIST*, vol. 58, no. 13, pp. 2126-2144, 2007.

[27] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. SIGIR '13.

[28] Paul Thomas and David Hawking, Evaluation by comparing result sets in context, CIKM' 06.

[29] Ryen W. White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. WWW '13.