Preference Based Evaluation Measures for Novelty and Diversity

Praveen Chandar and Ben Carterette
{pcr,carteret}@udel.edu
Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA 19716

ABSTRACT

Novel and diverse document ranking is an effective strategy that involves reducing redundancy in a ranked list to maximize the amount of novel and relevant information available to users. Evaluation for novelty and diversity typically involves an assessor judging each document for relevance against a set of pre-identified subtopics, which may be disambiguations of the query, facets of an information need, or nuggets of information. Alternately, when expressing a *preference* for document A or document B, users may implicitly take subtopics into account, but may also take into account other factors such as recency, readability, length, and so on, each of which may have more or less importance depending on user. A *user profile* contains information about the extent to which each factor, including subtopic relevance, plays a role in the user's preference for one document over another. A preference-based evaluation can then take this user profile information into account to better model utility to the space of users.

In this work, we propose an evaluation framework that not only can consider implicit factors but also handles differences in user preference due to varying underlying information need. Our proposed framework is based on the idea that a user scanning a ranked list from top to bottom and stopping at rank k gains some utility from every document that is relevant their information need. Thus, we model the expected utility of a ranked list by estimating the utility of a document at a given rank using preference judgments and define evaluation measures based on the same. We validate our framework by comparing it to existing measures such as α-nDCG, ERR-IA, and subtopic recall that require explicit subtopic judgments We show that our proposed measures correlate well with existing measures while having the potential to capture various other factors when real data is used. We also show that the proposed measures can easily handle relevance assessments against multiple user profiles, and that they are robust to noisy and incomplete judgments.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

ACM 978-1-4503-2034-4/13/07 ...$15.00.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]; H.3.4 [**Systems and Software**]: Performance Evaluation

Keywords: Novelty and Diversity, Evaluation

1. INTRODUCTION

The concept of relevance is the probably the most critical aspect of theoretical and practical information retrieval (IR) models. But which documents are relevant can differ from user to user depending on their exact information need, even if they start with the same keyword query. Queries can be ambiguous and/or underspecified, and the retrieval systems are required to handle these diverse information needs while providing novel information. Traditional IR evaluation also works under the assumption that documents are independently relevant separate from any user context The major drawback with this approach is that it does not penalize redundancy in rankings, potentially reducing the amount of novel information available to the user. Recently a subtopic based approach was introduced, to handle the redundancy problem and account for diverse information needs. The underlying information need for a query is decomposed into set of subtopics, and the number of novel subtopics that a document is relevant to (*i.e.* not seen earlier in the ranking) provides a measure of novelty. Various evaluation measures have been defined based on this approach [1, 9, 23, 27].

While subtopics are used to account for the diverse information needs of a query, the relation between them varies from user to user. For example, consider the query *living in India*. A person planning to visit India could be interested in *information for visitors and immigrants & how people live in India* whereas a student writing an essay would be more interested in *the history about life and culture in India.* Even though all of these subtopics seem relevant to the query, the importance of a subtopic is dependent on the user and the scenario in which the search was performed. It is well-known that user preferences are influenced not only by topical relevance but also by other factors such as readability, subtopic importance, completeness, etc. User profiles can be used to represent the combination of relevant subtopics and the above mentioned factors that precisely reflects the user's information need. Currently, there is no evaluation measure that (a) takes into account various factors affecting user preference, (b) handles multiple user profiles for a given query.

In this work, we propose an evaluation framework and metrics based on user preference for the novelty and diversity task. The framework revolves around the idea of assigning

utility scores that reflect each set of user's preference towards each document. The document utilities are estimated using a series of preference judgments collected conditional on previously ranked documents. Document utility at a given rank implicitly accounts for the subtopic coverage, novelty, topical relevance and the other factors as well. As pointed out earlier, the utility of document could differ for each user, thus user preference are obtained across a pool of users to account for diverse information need of a query. Evaluation metrics defined based on this framework directly models a user traversing a ranking from top to bottom seeking relevant and novel information for the issued query. Therefore, our proposed measures estimate the total utility of a ranked list available to the user for a given query.

The rest of the paper is organized as follows: a detail explanation of the existing evaluation framework and the existing metrics for novelty and diversity is provided in Section 2. We point out issues with the current method and propose preference-based evaluation measures in Section 3. A description of the datasets along with the experimental design employed in our work can be found in Section 4. We analyze in detail the performance of our metrics and compare it to various existing ones in Section 5. Finally, Section 6 summarizes our findings and sketches our future directions.

2. NOVELTY/DIVERSITY EVALUATION

Search result diversification is an effective strategy to deal with the diverse information needs of the user while reducing redundancy in the ranked list [19, 28, 25]. Several methods have been proposed to produce a ranking that maximizes relevance with respect to multiple information needs for a given query, starting with the maximum marginal relevance model of Carbonell et al. [4]. In addition to new models, the task demands new evaluation metrics, as traditional IR measures are focused on relevance with respect to a single user and do not penalize redundancy in results. Zhai et al. studied the subtopic retrieval task in the context of the TREC Interactive track [17], and defined simple evaluation measures such as subtopic recall and subtopic precision based on the relevance of documents to pre-defined subtopics. Clarke et al. proposed an evaluation strategy that decomposes underlying information needs of a query into information nuggets; document utility is determined by the number of novel nuggets covered by the document. NRBP, also introduced by Clarke et al. combines ideas from α-nDCG and Rank-Biased Precision [12]. Agarwal et al. focused on the diversity problem in the web domain by taking into account the importance of user intents via a probability distribution. Each of these measures will be described in more detail below.

Almost all of the existing measures are based on the idea of explicit *subtopics*: decompositions of a given query into several pieces of information (such as facets, intents, or nuggets) that account for various underlying information needs. In this framework, novelty is solely dependent on the document's relevance to a subtopic. System effectiveness is estimated by iterating over the ranked list, penalizing relevant documents relevant to subtopic(s) seen earlier in the ranking, and rewarding documents relevant to unseen subtopic(s).

2.1 Test Collection

Test collections such as those produced for the TREC Interactive tracks [17] and the TREC Question Answering tracks [26] consist of subtopic-level judgments in documents. The TREC Web track diversity datasets created to study the problem of novelty and diversity are most suitable to our work. These datasets comprise a set of topics, and for each topic a set of subtopics that were identified semi-automatically with the help of a tool that clusters reformulations of the given query. The tool combined evidences from clicks and reformulations to obtain clusters of queries; the track organizers used these clusters to manually pick the set of subtopics for a given target query.

Binary judgments of relevance were made by NIST assessors for each subtopic to each document. Note that the use of this method means that only subtopics evidenced by a large number of users will be present in the data; interpretations that are equally "real" yet less popular will not be represented when this method is used.

2.2 Evaluation Measures

Evaluation measures for novelty and diversity must account for both relevance and novelty in rankings. It is important that redundancy caused by documents containing previously retrieved information are penalized while documents containing novel information are rewarded; as described above, this is achieved using subtopic relevance judgments. A brief description of the commonly used metrics that employ a subtopic based approach is given below:

Subtopic recall measures the proportion of unique subtopics retrieved at a given rank [27]. Given that a query q has m subtopics, the subtopic recall at rank k is given by the ratio of number of unique subtopics contained by the subset of document up to rank k to the total number of subtopics m.

$$S\text{-}recall@k = \frac{\left|\bigcup_{i=1}^{k} subtopics(d_i)\right|}{m} \tag{1}$$

α-nDCG scores a result set by rewarding newly found subtopics and penalizing redundant subtopics [13]. Computation of the gain vector and a rank discount are key to α-nDCG. The gain vector is computed by summing over subtopics appearing in the document at rank i:

$$G[i] = \sum_{j=1}^{m} (1-\alpha)^{c_{j,i}-1} \tag{2}$$

where $c_{j,i}$ is the number of times subtopic j has appeared in documents up to (and including) rank i.

The most commonly used discount function is $\log_2(1+i)$, although other discount functions are possible. Summing gains over discounts gives *discounted cumulative gain*:

$$\alpha\text{-}DCG@k = \sum_{i=1}^{k} \frac{G[i]}{\log_2(1+i)} \tag{3}$$

α-DCG must be normalized to compare the scores against various topics. This is done by finding an "ideal" ranking that maximizes α-DCG, which can be done using a greedy algorithm. The ideal ranking computation is an NP-Complete problem [5]. The ratio of α-DCG to that ideal gives α-nDCG.

Intent-aware family Agrawal et al. [1] studied the problem of evaluating ambiguous web queries. They proposed evaluating a ranking against *each* subtopic (or "intent") by

any traditional IR measure, and then combining the results based on importance of subtopic. This gave rise to a family of measures that are known as *intent-aware*. Most traditional measures such as precision@k, average precision (AP), nDCG, etc. can be cast as intent-aware versions; for instance, intent-aware AP would be expressed as:

$$AP\text{-}IA = \sum_{i=1}^{m} P(i|q) AP_i \quad (4)$$

where m is the number of intents/subtopics, $P(i|q)$ is the probability that the user is interested in intent i for query q, and AP_i is average precision computed only with the documents relevant to intent i.

ERR-IA Expected Reciprocal Rank (ERR) is a measure based on "diminishing returns" for relevant documents [10]. According to this measure, the contribution of each document is based on the relevance of documents ranked above it. The discount function is therefore not just dependent on the rank but also on relevance of previously ranked documents.

$$ERR = \sum_{i=1}^{\infty} \frac{1}{i} R_i \prod_{j=1}^{i-1} (1 - R_j) \quad (5)$$

where R_i is a function of the relevance grade of the document at rank i (typically defined to be $(2^g - 1)/2^{g_m ax}$). ERR-IA is defined exactly as other intent-aware measures: a weighted average of ERR computed separately for each subtopic/intent [9]. We mention it separately because it has some appealing mathematical properties and it is one of the official measures of the TREC Web track [9].

D-Measure The D and the D# measures described by Sakai et al. [22] aim to combine two properties into a single evaluation measure. The first property is to retrieval documents covering as many intents as possible and second is to rank documents relevant to more popular intents higher than documents relevant to less popular intents.

3. PREFERENCE BASED FRAMEWORK

The subtopic-based evaluation framework focuses on estimating the effectiveness of a system based on topical and sub-topical relevance. In practice, there may be many other factors such as reading level, presentation, completeness, etc. that influence user preferences for one document over another in the context of novelty and diversity [8]. We could describe the information needs of a user that consists of various details, including specifics of pieces of information the user is interested in, reading level of the user, and so on in a *user profile*. Then we could view the goal of an evaluation measure as determining how well a ranking of documents satisfies a variety of user profiles.

In order to understand the concept of user profiles, let us consider an example query from the TREC Web track: *air travel information*. Table 1 shows the subtopics defined for the Web track's diversity task and provides the information needs of three different possible users for the given query (assuming we restrict ourselves to the TREC paradigm and represent the user's information need using only subtopics). We can think of user A as a first time air traveler looking for information on air travel tips and guidelines, user B as a journalist writing an article on the current quality of air travel and looking for statistics and reports to accomplish the task, and user C as an infrequent traveler looking restrictions and rules for check-in and carry-on luggages. Therefore, user A's profile for the above example query consists of subtopics d and e, user B's of c, and user C's of a and b. (In practice, the profiles would typically take into account other factors such as presentation, readability, and other factors as well, but none of this need be made explicit.)

Even if we restrict ourselves to modeling only subtopics, there are some issues with existing measures based on subtopics:

(a) subtopic identification is challenging and tricky as it is not easy to enumerate all possible information needs for a given query,

(b) measures often require many parameters to be set before use,

(c) measures assume subtopics to be independent of each other but in reality this is not true.

Let us refer to Table 1 to consider these issues. First, given the granularity of these subtopics, it would not be difficult to come up with additional subtopics that are not in the data. Top-ranked results from a major search engine suggest subtopics such as "Are airports currently experiencing a high level of delays and cancellations?", "I am disabled and require special consideration for air travel; help me find tips.", and "My children are flying alone, I am looking for tips on how to help them feel comfortable and safe." Are users with these needs going to be satisfied by a system that optimizes for the limited set provided?

Second, measures like α-nDCG and ERR-IA have a substantial number of parameters that must be decided on. Some are explicit, such as α (the penalization for redundancy) [15] or $P(i|q)$ (the probability of an intent/subtopic given a query[1]). Others are implicit, hidden in plain sight because they have "standard" settings: the log discount of α-nDCG or the grade value R_i of ERR-IA, for instance. Each of these parameters requires some value; it is all too easy to fall back on defaults even when they are not appropriate.

Third, some subtopics are clearly more related to each other than others (in fact, we used this similarity to create the profiles). Documents that are relevant to subtopic c are highly unlikely to also be relevant to any of the other subtopics, but it is more likely that there are pages relevant to both subtopics a and b.

In this work, we sidestep these issues by proposing an evaluation framework that simply allows users to express preferences between documents. Their preferences may be based on topical or subtopic relevance, but they may also be based on any other factors that are important to them. Preferences can be obtained over many users to capture the varying importance of topics and factors, and when a sufficiently large set of preferences has been obtained, systems can be evaluated according to how well they satisfy those users. Preference judgments have only scantly been used in IR evaluation, having been introduced by Rorvig [20] but not subject to empirical study until recently [7, 2]. Comparison studies between absolute and preference judgments show that preference judgments can often be made faster than graded judgments, with better agreement between assessors (and more consistency with individual assessors) [7] while making much finer distinctions between documents.

[1]The original definition of α-nDCG has parameters for subtopic weights as well.

	subtopic	*user A*	*user B*	*user C*
a.	What restrictions are there for checked baggage during air travel?			✓
b.	What are the rules for liquids in carry-on luggage?			✓
c.	Find sites that collect statistics and reports about airports		✓	
d.	Find the AAA's website with air travel tips.	✓		
e.	Find the website at the Transportation Security Administration (TSA) that offers air travel tips.	✓		

Table 1: An example topic (*air travel information*) along with its subtopics from the TREC Diversity dataset and three possible user profiles indicating the interests of different users.

Chandar and Carterette [8] introduced a preference-based framework similar to ours, but there exists no evaluation measure that incorporates preference judgments directly for novelty and diversity. Moreover, that work focused only on ranking novel documents, without considering the more general question of diversity—that different users will have different preferences depending on their profile.

3.1 Test Collection

Chandar and Carterette's preference-based framework is based on so-called *levels* of preference judgments. We use a similar idea; in this work, a test collection of preferences for novelty and diversity consists of two different types of preference judgments:

1. simple pairwise preference judgments, in which a user is shown two documents and asked which they prefer.
2. *conditional* preference judgments, in which a user is shown three or more documents and asked to express a preference between two of them, supposing they had read the others.

Simple pairwise preferences produce a relevance ranking: given a pair of documents, assessors select the preferred document based on some criteria. We expect topical relevance to be the primary criteria, although many criteria (such as ease of reading, completeness of information, salience of article, etc.) could factor into an assessor's choice. Since different users may have different needs and different preferences for the same query, pairs can be shown to multiple assessors to get multiple preferences. Over a large space of assessors, we would expect that documents are preferred proportionally according to the relative importance of the subtopics they are relevant to, with various other factors influencing finer-grained orderings.

Simple pairwise preferences cannot capture novelty; in fact, two identical documents should be equally preferred in all pairs in which they appear and therefore end up tied in the final ordering. Conditional preference judgments attempt to resolve this by asking for a preference for a given pair of document *conditional on* the information in other documents shown to the assessor at the same time. The assessor is asked to read those documents, then select which of the remaining two they would like to see *next*.

Figure 1 illustrates conditional preferences with a triplet of documents: the assessor would read document X, then select which of A or B they would like to see next[2] We

[2]Note that any document may be placed at the top of a triplet; it need not be the most preferred document among the simple pairwise preferences.

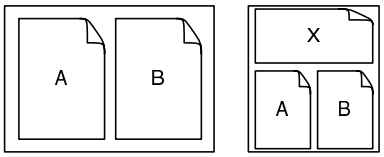

Figure 1: Left: a simple pairwise preference for which an assessor chooses A or B. Right: a triplet of documents for conditional preference judgments. An assessor would be asked to choose A or B conditional on having read X.

expect the assessor's choice to be based not only on topical relevance, but also on the amount of *new* information given what is provided in the top document. Again, they can use other factors in their preferences, but novelty should be a primary consideration: if X is identical to A, we expect them to choose B, and then a system that ranks X and A adjacent would be penalized for failing to rank B after X.

Similarly, we could obtain preferences with quadruplets of documents, quintuplets of documents, and so on. In practice it becomes increasingly difficult for assessors to make such fine distinctions, so we limit to only obtaining judgments on triplets. A triplet in our framework corresponds to Chandar and Carterette's "level 2" judgments; as they showed, these judgments capture most of the necessary information about novelty. Preferences conditional on greater numbers of other documents contribute less and less [8].

3.2 Preference-Based Evaluation Measure

We propose a model-based measure using preferences to assess the effectiveness of systems for the novelty and diversity task. Model based measures can be composed from three underlying models: *browsing model*, *document utility*, and *utility accumulation* [6]. The way users interact with the ranked list is defined by the browsing model; we rely on the most accepted model in which the user scans documents down a ranked list one-by-one and stops at some rank k. The document utility model defines the amount of utility provided by a single document, and utility accumulation models the total utility derived during browsing.

We define our utility based model for novel and diversity ranking task as follows: a user scanning documents down a ranked list derives some utility $U(d)$ from each document and stops at some rank k. We hypothesize that the utility of a document at rank i is dependent on previously ranked document (*i.e.* d_1 to d_{i-1}). Given a probability distribution for a user stoping at rank k, the utility accumulation model

can be defined as:

$$Prf = \sum_{k=1}^{n} P(k)U(d_1, ..., d_k) \quad (6)$$

where $P(k)$ is the probability that a user stops at rank k and $U(d_1, ..., d_k)$ is the total utility of the documents from ranks 1 through k.

We simplify this by formulating $U(d_1, ..., d_k)$ as a sum of individual document utilities conditional on documents ranked before:

$$Prf = \sum_{k=1}^{n} P(k) \sum_{i=1}^{k} U(d_i|S) \quad (7)$$

where $P(k)$ is the probability that a user stops at rank k, $U(d_i|S)$ gives the utility of the document at rank i conditional on a set of previously ranked document S, and the sum from $i = 1$ to k gives the total utility of all documents from ranks 1 through k. There are two main components in the above equation: the probability that a user stops at a given rank ($P(k)$) and the utility of a document conditioned of previously ranked documents ($U(d_i|S)$). Carterette demonstrated different ways to model the stopping rank from the various ad-hoc measure such as Rank Biased Precision [16], nDCG, and Reciprocal Rank [6].

1. $P_{RBP}(k) = (1-\theta)^{k-1}\theta$
2. $P_{DCG}(k) = \frac{1}{log(k+1)} - \frac{1}{log(k+2)}$
3. $P_{RR}(k) = \frac{1}{k(k+1)}$

Finally, we define the document utility model in which the document utility at a given rank is conditioned on previously ranked documents. The utility of the document at rank i is given by $U(d_i)$ for $i = 1$ since at rank 1 the user would not have seen any other documents and therefore would not be conditioning on any other documents. For subsequent ranks, utility is $U(d_i|d_{i-1}, ...d_1)$, indicating that the utility depends on documents already viewed.

Now our goal is to estimate these utilities using preference judgments. Since we have simple pairwise preferences and conditional preferences in triplets, we decompose the document utility model as follows:

$$U(d_i|S) = \begin{cases} U(d_i), & \text{if } i \text{ is 1} \\ U(d_i|d_{i-1}), & \text{if } i \text{ is 2} \\ F(\{U(d_i|d_j)\}_{j=1}^{i-1}), & \text{if } i > 2 \end{cases} \quad (8)$$

where the function $F()$ takes an array of conditional utilities ($U(d_i|d_j)$).

The utility $U(d_i)$ can be directly obtained using the pairwise judgments; we simply compute it as the ratio of number of times a document was preferred to the number of times it appeared in a pair. The utilities $U(d_i|d_{i-1})$ can similarly be obtained from the conditional preferences, computed as the ratio of the number of times d_i was preferred conditional on d_{i-1} appearing as the "given" document to the number of times it appear with d_{i-1} as the "given" document. Note that these utilities can be computed regardless of how many times a document has been seen, how many different assessors have seen it, how much disagreement there is between assessors, and so on. Although, a document must be shown at least few time in order to determine its relevance estimate. An estimate of the document's utility is obtain using the ratio of number of times the document was preferred to the number of time it was shown.

We experiment with two functions for $F()$: *average* and *minimum*. The intuition behind these functions can be explained with the help of an example. Consider a ranking $R = \{d_1, d_2, d_3\}$. According to equation 8 the utility of d_3 depends on $U(d_3|d_1)$ and $U(d_3|d_2)$. The minimum function assumes that d_3 cannot be any more useful conditional on both d_1 *and* d_2 than it is on either one separately, thus giving a sort of worst-case scenario. The average function assumes that the utility of d_3 conditional on both d_1 and d_2 is somewhere in between its utility conditioned on each separately, giving d_3 some benefit of the doubt that it may contribute something more when appearing after both d_1 and d_2 than it does when appearing after either one on its own.

Our measure as defined is computed over the entire ranked list. In practice, measures are often computed only to rank 5, 10, or 20 (partially because relevance judgments may not be available deeper than that). When we compute the measure to a shallower depth, we must normalize it so that it will average over a set of queries. As a final step in the computation of $nPrf$, we normalize equation 7 cut off at rank K by the ideal utility score.

$$nPrf[K] = \frac{Prf[K]}{I\text{-}Prf[K]} \quad (9)$$

where $I\text{-}Prf[K]$ is the ideal utility score that could be obtained at rank K. This can be obtained by selecting the document with the highest utility value conditioned on previously ranked documents. Document (d_1) with the highest utility value takes rank 1 and the document with highest utility when conditioned on d_1 takes rank 2 and so on.

Table 2 provides an example showing the distinction between our preference based measure and α-$nDCG$ based on the user profiles in Table 1. The document utilities are estimated by obtaining the preference judgements for all documents from all three users. We would expect the users' preferences to be consistent with their information need, for example user A would prefer d_1 and d_2 consistently to other documents that are not relevant to their needs (but relevant to other needs). Notice that α-$nDCG$ weighs all subtopics equally but the preference measure takes into account the dependency between the subtopics.

4. EXPERIMENT DESIGN

In Section 3.2, we proposed various evaluation measures based on a user model for novelty and diversity. Evaluation of the proposed metrics is challenging since there is no ground truth to compare to; there are only other measures. Approaches used in the past to validate newly introduced metrics include comparing the proposed measure to existing measures or click metrics [18, 11]; using user preferences to compare the metrics [24]; and evaluating the metric on various properties such as discriminative power [21]. While each of these approaches have their own advantages, we argue that comparison of existing measures to our measures using simulated data is suitable for this work.

Remember, our goal is to build evaluation measures for our preference based framework that assigns utility scores to a document based on user preferences. In reality, user preferences are based on various implicit factors that include

	documents	subtopics a	b	c	d	e
user A	d_1	✓				
	d_2		✓			
user B	d_3			✓		
	d_4			✓		
user C	d_5				✓	
	d_6					✓

List1	List2	
d_1	d_1	
d_2	d_3	
d_3	d_5	
1.0	1.0	α-$nDCG$
0.9	1.0	Preference Measure

Table 2: Synthetic example with 6 documents and 5 subtopics. The first ranked list does not satisfy all users where as the second one does but both rankings are scored by equally by α-nDCG, while the preference metrics are able to distinguish the difference.

subtopic relevance as well as many other properties. Since prior work [8] has suggested that presence of subtopics in a document plays a major role in user preferences, we believe it is important to validate our measures when user preferences are based solely on subtopic information. We therefore rely on the existing data with subtopic information to *simulate* user preferences.

4.1 Data

In our experiments, we used the ClueWeb09 dataset[3] consisting of one billion web pages (5 TB compressed, 25 TB uncompressed), in ten languages, crawled in January and February 2009. A subset of this collection with only English documents was used for the diversity task at TREC in 2009/10/11 [14]. A total of 150 queries have been developed and judged for the TREC Web track; the number of subtopics for each ranges from 3 to 8. For the diversity task, subtopic level judgments are available for each subtopic indicating the relevance of a document to each subtopic along with the general topical relevance. We also acquired the experimental runs submitted to TREC each year by Web track participants. A total of 48 systems were submitted by 18 groups in 2009, 32 system by 12 groups in 2010, and 62 systems by 16 groups in 2011.

4.2 Simulation of Users and Preferences

In order to verify and compare our metrics against existing measures, we acquire preferences by simulating them from subtopic relevance information. These will be based on the preferences of simulated users that are modeled by groupings of subtopics (as in Table 1). In this way we use only data that is provided as part of the TREC collection, and therefore achieve the fairest and most reproducible possible comparison between evaluation measures. In reality, our measure is well-suited for crowd-sourced assessments in a way that other measures are not, but we save that experiment for future work.

We created our user profiles by generating search scenarios for each query and marking subtopics relevant to the scenario. In Section 3, we explained our reasoning behind the user profiles in Table 1 for the query *air travel information*; we use the same approach to obtain the user profiles for all TREC queries. The user profiles were created by the authors of this paper and have been made available for public download at `http://ir.cis.udel.edu/~ravichan/data/profiles.tar`. In addition, there is a mega-user that we refer to as the "TREC profile"; this user is equally interested in all subtopics.

[3]http://lemurproject.org/clueweb09.php

These profiles are used to determine the outcome of preferences. For simple pairwise preferences, we always prefer the document with greater number of subtopics relevant to the user profile. In the case of a tie, we make a random choice between the left or right document. For conditional preferences, we have three documents (left, right, and top); between the left and the right, we prefer the document that contains the greater number of subtopics relevant to the user profile and not present in the top document. Preference judgments obtained this way are used to compute our preference measure. Finally, using the "TREC profile" to simulate preferences for our measure offers the most direct comparison to other measures.

5. ANALYSIS

We have presented a family of preference-based measures for evaluating systems based on novelty and diversity, and outlined the advantages of our metrics over existing subtopic-based measures. In this section, we demonstrate how our metrics take into account the presence of subtopics implicitly by comparing them with α-nDCG, ERR-IA, and s-recall.

5.1 System Ranking Comparisons

5.1.1 System Performance

We evaluated all experimental runs submitted to TREC in 2009, 2010, and 2011 using our proposed measure with three different stopping probabilities $P(k)$ and two different utility aggregation functions $F()$. Figure 2 shows the performance of systems with respect to both α-nDCG and our preference measure computed with $P_{RBP}(k)$ and $F_{avg}()$ functions and preferences simulated using the "TREC profile". Each point represents a TREC participant system; they are ordered on the x-axis by α-nDCG. Black circles give α-nDCG values as computed by the `ndeval` utility used for the Web track; blue x's indicate the preference measure score for the same system. In these figures we can see that the preference measure is roughly on the same scale as α-nDCG, though typically $0.1 - 0.2$ lower in an absolute sense.

Each increase or drop in the position of x's indicates disagreement with α-nDCG. The increasing trend of the curves in Figure 2 indicates that the correlation between the preference measure and α-nDCG is high. A similar trend was observed while using different $P(k)$ and $F()$ functions as well (not shown). Both α-nDCG and our preference measure agree on the top ranked system in 2009 and 2010.

We analyzed the reason behind disagreement by carefully looking at the actual ranked lists. We investigated how α-nDCG and our proposed measures reward diversified sys-

	ERR-IA@20	s-recall@20
α-nDCG@20	0.893	0.828
ERRIA@20	-	0.739

Table 3: Kendall's τ correlation values between the existing evaluation measures. Values were computed using 48 submitted runs in TREC 2009 dataset.

tems on a per topic basis. Based on our analysis, the major reason for disagreement is that α-nDCG penalizes systems that miss documents containing many unique subtopics more harshly than the preference measure does. Much of the variance in α-nDCG scores is due to differences in rank position of the documents with the greatest number of unique subtopics. In practice, this explains the lower scores returned by the preference measure as well.

5.1.2 Rank Correlation Between Measures

We measure the stability of our metrics using Kendall's τ by ranking the experimental runs under different effectiveness measures. Kendall's τ ranges from -1 (lists are reversed) to 1 (lists are exactly the same), with 0 indicating essentially a random reordering. Prior work suggest that a τ value of 0.9 or higher between a pair of rankings indicates high similarity between rankings while a value of 0.8 or lower indicates significant difference [3].

Figure 3 summarizes the rank correlations between existing subtopic-based metrics and our proposed preference metric using all three $P(k)$ (plus using no $P(k)$ at all—equivalent to a uniform stopping probability) and both $F()$ functions, simulating preferences with the "TREC profile". The correlations are fairly high across TREC datasets, $P(k)$ functions, and $F()$ functions. The $P_{DCG}(k)$ rank function fares worst, with correlations dipping quite a bit for the 2010 data in particular. Subtopic recall is a very simple non-rank based metric for diversity and thus the Kendall's τ values are expected to be slightly lower.

For comparison, Table 3 shows the Kendall's τ correlation values between α-nDCG, ERR-IA and s-recall. These correlations are similar to those in Figure 3, suggesting that the ranking of systems given by our preference measure varies no more than the rankings of systems given by any two standard measures.

There is almost no difference between the correlations for $F_{avg}()$ and $F_{min}()$ functions for aggregating utility. In fact, the correlation between preference measures computed with those two is nearly 1. Thus we can conclude that the choice of $F()$ (between those two options) does not matter. There is a great deal of difference depending on choice of $P(k)$, however, and thus this is a decision that should be made carefully based on the observed behavior of users.

5.2 Evaluating Multiple User Profiles

The experiments above are based on the "TREC profile", a user profile that considers every subtopic to be equally relevant. In this experiment, we demonstrate the ability of our methods to handle multiple, more realistic user profiles and show the stability of our metrics. Measures based on absolute subtopic judgments cannot naturally incorporate multiply-judged documents. One must average judgments, or take a majority vote, or use some other scheme. In contrast, judgments from multiple users can be incorporated easily into our preference framework in the estimation of document utilities, as the document utility is simply the ratio of number of times a document was preferred to the number of times it appeared in a pair, regardless of which user or assessor happened to see it.

We simulate preferences for each of our user profiles for each topic in the TREC set. We compute the preference measure using each profile's preferences separately (giving at least three separate values for each system: one for each user profile), and then use the full set of preferences obtained to compute a single value of the measure. Note that the latter case is *not* the same as computing the preference measure with the "TREC profile": the TREC profile user uses all subtopics to determine the outcome of a preference, while individual users would never use a subtopic that is not relevant to them to determine the outcome of a preference.

We can also compute subtopic-based measures such as α-nDCG against our profiles. To do this, we simply assume that only the subtopics that are relevant to the profile "count" in the measure computation. We will compare values of measures computed this way to our preference measures.

Our hypothesis for this experiment is twofold: 1) that the preference measure computed for a single profile will correlate well to subtopic-based measures computed against the same profile; 2) that the preference measure computed with preferences from all profiles will *not* be the same as an average of the individual profile measures, and also not the same as subtopic-based measures computed as usual. In other words, that the preference measure based on preferences from many different users is measuring something *different* than the preference measure based on preferences from one user, and also different from the subtopic measures.

Figure 4 shows the results of evaluating systems using user profile 1, 2, and 3 for each topic and averaging over topics (note that the user profile number is arbitrary; there is nothing connecting user profile 1 for topic 100 to user profile 1 for topic 110). We can see that the system ranking changes for both α-nDCG and the preference measure, as expected. The correlation between the two remains high: $0.83, 0.88$, and 0.82 for user profile 1, 2, and 3 respectively. This is in the same range of correlation values that we saw in Figure 3, and supports the first part of our hypothesis.

Figure 5 shows the results of evaluating systems with all user profiles, comparing to the evaluation with the TREC profile and with α-nDCG computed with all subtopics. Note here that all three rankings are different, as evidenced by the τ correlations reported in the inset tables. This supports the second part of our hypothesis: that allowing many different users the opportunity to express their preferences can result in a different ranking of systems than treating all assessors as equivalent, as the TREC profile and α-nDCG do.

5.3 Incomplete Judgments

The test collection procedure discussed in Section 3.1 requires two sets of judgments: pairwise and conditional preferences. The number of pairwise judgments increases quadratically with increase in number of documents in the pool; it is not feasible to collect a complete set of preferences. We envision that our measure would *always* be computed with incomplete judgments. For this experiment we test the stability of our measures by comparing the system rankings obtained by using all preference judgments against a set of incomplete judgments.

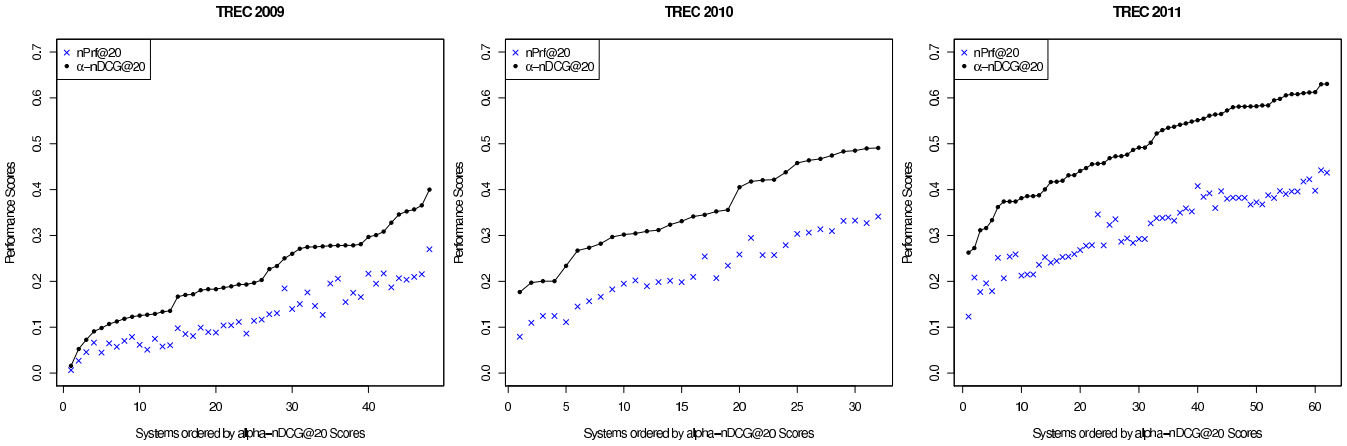

Figure 2: TREC 09/10/11 diversity runs evaluated with our preference based metric at rank 20 (nPrf@20) with P_{RBP} and $F_{Average}$. Compare to α-nDCG scores.

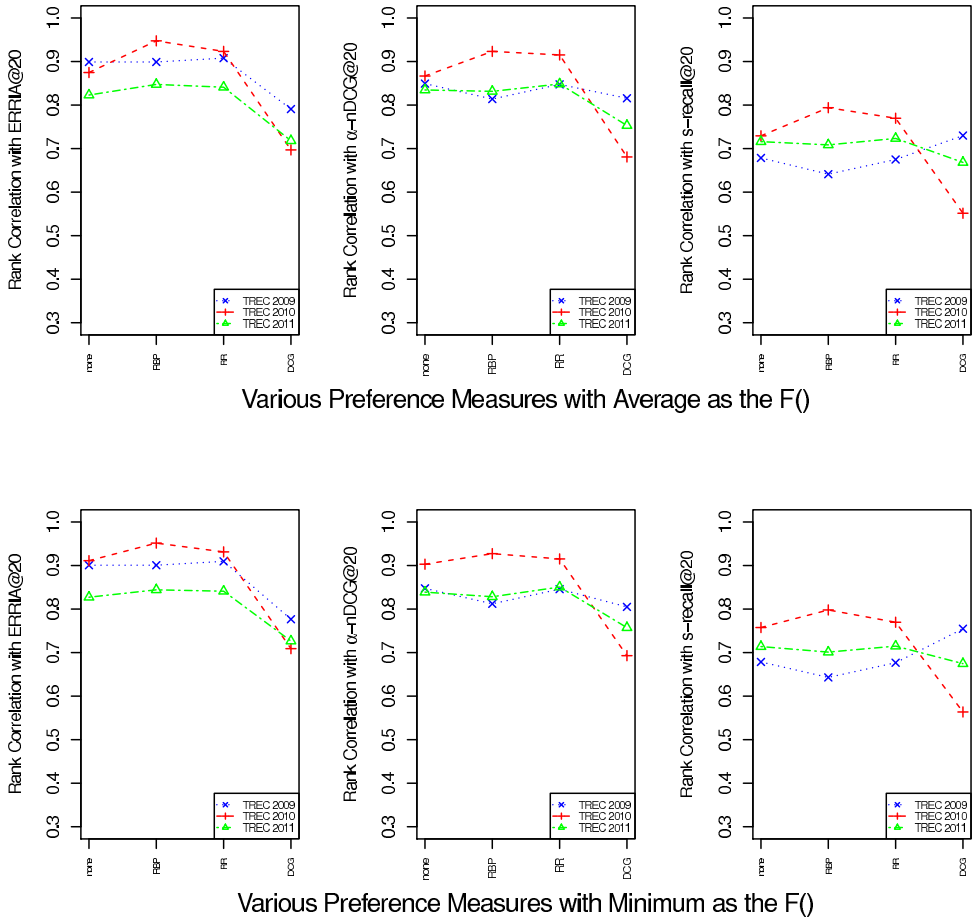

Figure 3: Kendall's τ correlation values between our proposed measures and α-nDCG, ERR-IA, s-recall. Values were computed using the submitted runs in the TREC 2009/10/11 dataset. The scores for various $P(k)$ and F() are shown.

To do this, we randomly select N triplets of documents for each query. For each triplet, one document is randomly selected to be the "top" document that the other two would be judged conditional on. Though we do not explicitly obtain simple pairwise preferences, we expect that there will be enough cases in which the top document is not relevant to the user profile that they must fall back on a simple pairwise comparison. We then sample 5 user profiles (with replacement) from those defined for the topic and simulate their preferences for the triplet. In this way we obtain $5N$ prefer-

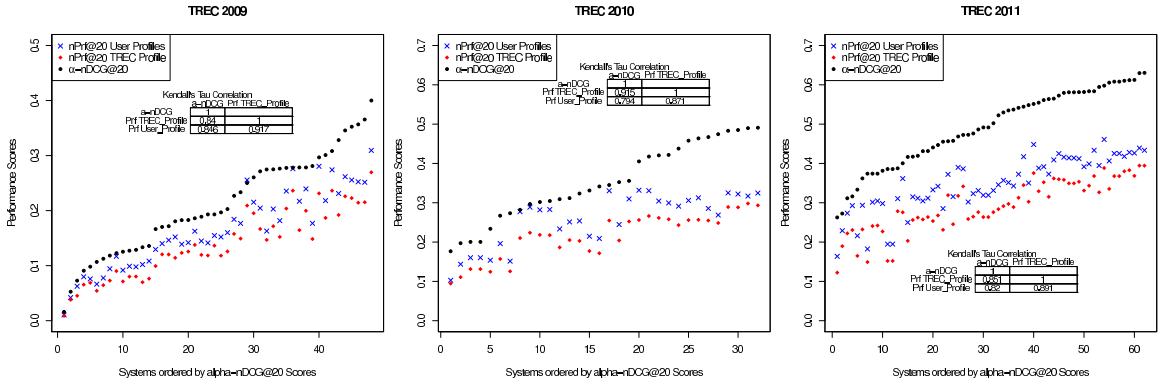

Figure 5: Comparison between α-nDCG, our preference measure computed using the TREC profile, and our preference measure computed using a mix of user profiles. Note that all three rankings, while similar, have substantial differences as well.

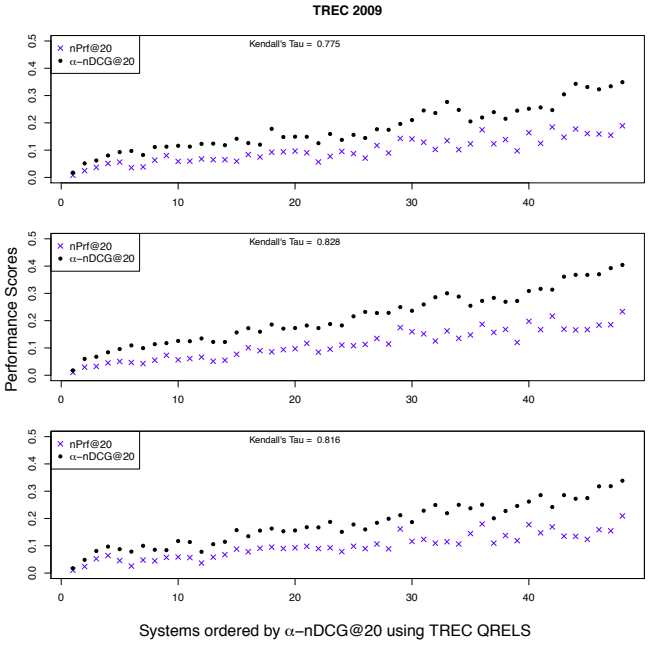

Figure 4: Comparison between α-nDCG and our preference measure computed against user profiles 1 (top), 2 (middle), and 3 (bottom) for TREC 2009 systems.

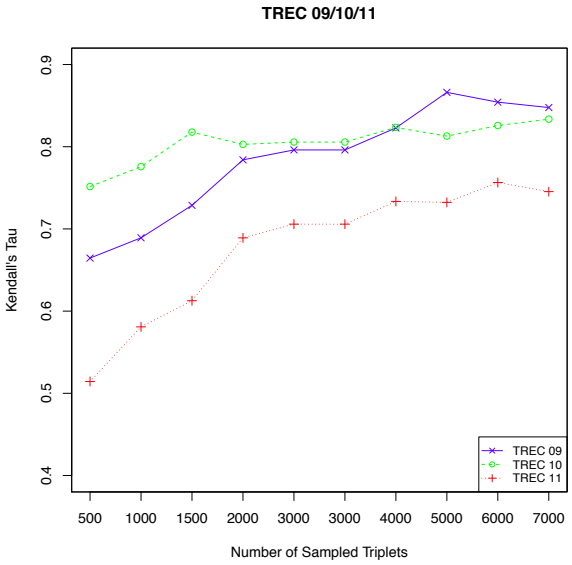

Figure 6: TREC 09/10/11 diversity runs evaluated with our preference based metric at rank 20 (nPrf@20) with P_{RR} and $F_{Minimum}$ using single assessor with complete judgments and multiple assessor with incomplete judgments.

ences for each topic in a similar way as would be done in a real crowd-sourced assessment. We use those preferences to compute our measure, then compute the correlation to the measure computed with all available preferences. We repeat this 10 times for each topic, measure the correlation each time, and average the correlations.

Figure 6 shows the correlation between the system rankings when evaluated using complete judgements and increasing numbers of preferences. Correlation tends to increase as the number of preferences increases, though it does not reach 0.9. This may be partly because user profiles are not evenly represented in the preferences (which is in fact more realistic than when they are, as in the full-preference case), and partly because our max number of preferences is still a fairly small fraction of the total number possible: even selecting triplets from only 100 documents, there are over 161,000 possible triplets, of which we have only obtained less than 5%! Thus we expect that continuing to increase the number of triplets would continue to push the correlations higher, even though we see dips in the trend (due to variance).

6. CONCLUSION AND FUTURE WORK

In this work, we proposed a novel evaluation framework and a family of measures for IR evaluation. Our measure incorporates novelty and diversity, but can also incorporate any property that influences user preferences for one document over another. Our measure is motivated directly by

a user model and has several advantage over the existing measures based on explicit subtopic judgments: it captures subtopics implicitly and at finer-grained levels, it accounts for subtopic importance and dependence as expressed by user preferences, and it requires few parameters—only a stopping probability function, for which there are several well-accepted options that can be chosen from by comparing to user log data. It correlates well with existing measures, but also clearly measures something different (which is a positive for a new measure).

This framework and measure is most well-suited for assessments done by crowd-sourcing. In a crowd-sourced assessment, we would naturally have a large user base with a wide range of preferences. Over a large number of preferences, the most important subtopics and intents would naturally emerge; documents relevant to those would become the documents with the highest utility scores. Yet the conditional judgments would prevent too many documents with those subtopics from reaching the top of the ranking. The measure is designed to handle multiple judgments, disagreements in preferences, *and* novelty of information, and as such it is novel to the information retrieval literature.

The clearest direction for future work is to perform an actual crowd-sourced assessment and determine whether our preference measure correlates better with human judgments of system performance than other measures. We plan to start this immediately. Another direction for future work is using triplets in a learning-to-rank algorithm to learn a novelty ranker. Since many learning algorithms are based on pairwise preferences, it seems a natural extension to triplets.

Acknowledgments: This work was supported in part by the National Science Foundation (NSF) under grant number IIS-1017026. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. *Proceedings of WSDM '09*, page 5, 2009.

[2] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In *Proceedings of CIKM '11*, page 201, New York, USA, 2011. ACM Press.

[3] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR '04*, page 25, New York, USA, 2004. ACM Press.

[4] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of SIGIR '98*, pages 335–336, 1998.

[5] B. Carterette. An analysis of np-completeness in novelty and diversity ranking. *Information Retrieval*, 14(1):89–106, Dec. 2010.

[6] B. Carterette. System effectiveness, user models, and user utility. In *Proceedings of SIGIR '11*, page 903, New York, USA, 2011. ACM Press.

[7] B. Carterette, P. N. Bennett, D. M. Chickering, and T. Susan. Here or there preference judgments for relevance. In *Proceedings of ECIR '08*, pages 16–27, 2008.

[8] P. Chandar and B. Carterette. Using preference judgments for novel document retrieval. *Proceedings of SIGIR '12*, page 861, 2012.

[9] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, May 2011.

[10] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. *Proceeding of CIKM '09*, page 621, 2009.

[11] C. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of WSDM '11*, pages 75–84. ACM, 2011.

[12] C. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. *Advances in Information Retrieval Theory*, pages 188–199, 2010.

[13] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. *Proceedings of SIGIR '08*, page 659, 2008.

[14] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the trec 2011 web track. In *Proceedings of The Eighteenth Text REtrieval Conference TREC*, pages 1–9, Gaithersburg, Maryland, 2011. NIST.

[15] T. Leelanupab, G. Zuccon, and J. M. Jose. A query-basis approach to parametrizing novelty-biased cumulative gain. In *Proceedings of the Third international conference on Advances in information retrieval theory*, ICTIR '11, pages 327–331. Springer-Verlag, 2011.

[16] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, Dec. 2008.

[17] P. Over. Trec-6 interactive track report. In *The Sixth Text Retrieval Conference (TREC-6)*, pages 57–64, 1998.

[18] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of CIKM '08*, pages 43–52, New York, USA, 2008. ACM.

[19] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of WWW Õ10*, pages 781–790, New York, USA, Apr. 2010. ACM.

[20] M. E. Rorvig. The simple scalability of documents. *Journal of the American Society for Information Science*, 41(8):590–598, Dec. 1990.

[21] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of SIGIR '06*, pages 525–532, New York, USA, 2006. ACM.

[22] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Lin. Simple evaluation metrics for diversified search results. In *Proceedings of the 3rd International Workshop on Evaluating Information Access (EVIA)*, 2010.

[23] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of SIGIR '11*, pages 1043–1052. ACM, 2011.

[24] M. Sanderson, M. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceeding of SIGIR '10*, pages 555–562. ACM, 2010.

[25] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. *Proceedings of WWW '10*, page 881, 2010.

[26] E. M. Voorhees and H. T. Dang. Overview of the trec 2005 question answering track. In *In TREC 2005*, 1999.

[27] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR '03*, pages 10–17. ACM, 2003.

[28] W. Zheng, X. Wang, H. Fang, and H. Cheng. Coverage-based search result diversification. *Information Retrieval*, 2011.