Simple Evaluation Metrics for Diversified Search Results

Tetsuya Sakai[†] Nick Craswell[‡] Ruihua Song[†]
Stephen Robertson[*] Zhicheng Dou[†] Chin-Yew Lin[†]
[†]Microsoft Research Asia, PRC [‡]Microsoft, USA [*]Microsoft Research Cambridge, UK
tetsuyasakai@acm.org

ABSTRACT

Traditional information retrieval research has mostly focussed on satisfying clearly specified information needs. However, in reality, queries are often ambiguous and/or underspecified. In light of this, evaluating search result diversity is beginning to receive attention. We propose simple evaluation metrics for diversified Web search results. Our presumptions are that one or more interpretations (or intents) are possible for each given query, and that graded relevance assessments are available for intent-document pairs (as opposed to query-document pairs). Our goals are (a) to retrieve documents that cover as many intents as possible; and (b) to rank documents that are highly relevant to more popular intents higher than those that are marginally relevant to less popular intents. Unlike the *Intent-Aware* (IA) metrics proposed by Agrawal *et al.*, our metrics successfully avoid ignoring minor intents. Unlike α-nDCG proposed by Clarke *et al.*, our metrics can accomodate (i) which intents are more likely than others for a given query; and (ii) graded relevance within each intent. Furthermore, unlike these existing metrics, our metrics do not require approximation, and they range between 0 and 1. Experiments with the binary-relevance Diversity Task data from the TREC 2009 Web Track suggest that our metrics corrrelate well with existing metrics but can be more intuitive. Hence, we argue that our metrics are suitable for diversity evaluation given either the intent likelihood information or per-intent graded relevance, or preferably both.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

ambiguity, diversity, evaluation, novelty, test collection

1. INTRODUCTION

Traditional information retrieval research has mostly focussed on satisfying clearly specified information needs. However, in reality, queries are often ambiguous and/or underspecified [7]. When the retrieval system has no or little knowledge of the user, the best it can do is to produce output that reflect several *interpretations* (or *intents*) of such queries. In light of this, evaluating search result diversity is beginning to receive attention [1, 6, 5, 11, 18, 20]. We are particularly interested in evaluating diversified Web search engine results, where each result is assumed to be a single ranked list of documents. While richer forms of output such as those involving document clustering and aggregating results from different media and information sources are possible today, a flat ranked list remains a simple and effective method for presenting retrieved material to the user.

We propose simple evaluation metrics for diversified Web search results. Our presumptions are that one or more interpretations (or intents) are possible for each given query, and that graded relevance assessments are available for intent-document pairs (as opposed to query-document pairs). Our goals are (a) to retrieve documents that cover as many intents as possible; and (b) to rank documents that are highly relevant to more popular intents higher than those that are marginally relevant to less popular intents. Our approach is to evaluate Properties (a) and (b) separately first, and then combine the outcome.

Unlike the *Intent-Aware* (IA) metrics proposed by Agrawal *et al.* [1] our metrics successfully avoid ignoring minor intents. Unlike α-nDCG proposed by Clarke *et al.* [6], our metrics can accomodate (i) which intents are more likely than others for a given query; and (ii) graded relevance within each intent. Furthermore, unlike these existing metrics, our metrics do not require approximation, and are guaranteed to range between 0 and 1 and are therefore suitable for comparison and averaging across topics. Experiments with the binary-relevance Diversity Task data from the TREC 2009 Web Track suggest that our metrics corrrelate well with existing metrics but can be more intuitive. Hence, we argue that our metrics are suitable for diversity evaluation given either the intent likelihood information or per-intent graded relevance, or preferably both. We expect that these two types of information will become available with diversity test collections in the very near future because (1) methods exist to estimate the likelihood of each intent [1, 19]; and (2) search engine companies already use graded relevance for evaluation, and it is natural for them to adopt graded relevance also for multiple-intent queries.

The remainder of this paper is organised as follows. Section 2 describes previous work related to the present study, with a focus on evaluation metrics. Section 3 proposes new metrics for evaluating diversified search results, and clarifies their advantages over exising metrics. Section 4 describes an

EVIA 2010 June 15, Tokyo, Japan.

experiment using the Diversity Task runs from the TREC 2009 Web Track to support our claims. Section 5 concludes this paper.

2. RELATED WORK

2.1 S-recall

Zhai, Cohen and Lafferty [22] proposed a simple metric called *S-recall* (subtopic recall) for evaluating *subtopic retrieval*. Suppose that a topic consists of n subtopics, and that a document can either be relevant or nonrelevant to each subtopic. Let d_r denote the document at rank r, and let $I(d_r)$ denote the set of subtopics to which d_r is relevant. Then, S-recall at document cutoff l is defined as $|\bigcup_{r=1}^{l} I(d_r)|/n$, i.e. the proportion of subtopics that are covered by the top l documents. When used on its own, it is a rather crude metric, since it disregards the positions of relevant documents within top l and does not accomodate graded relevance. The only difference between S-recall and *instance recall* used earlier at the TREC interactive track [10] appears to be that the former is used in ranked retrieval while the latter was used in set retrieval.

Based on S-recall, Zhai, Cohen and Jafferty further defined S-precision and WS-precision at a given S-recall level [22]. Both of them involve computation of an NP-hard problem, and therefore approximation is required.

Carterette and Chandar [2] used S-recall for a task similar to subtopic retrieval, which they called *faceted topic retrieval*. They argued that for their task, S-recall at l_{min}, where l_{min} is the minimum rank at which perfect S-recall can be achieved, is the most natural evaluation measure. However, they also point out that finding l_{min} is NP-hard.

Our proposed metrics for evaluating diversified results use S-recall as one of its components. However, our methods do not involve NP-hard problems.

2.2 Intent-Aware Metrics

Agrawal *et al.* [1] proposed a family of *Intent-Aware* (IA) metrics, and primarily examined nDCG-IA. Let i be an intent (called "category" in [1]) and suppose that for each query q, the probabilities of different intents $P(i|q)$ are given. Then nDCG-IA for q at cutoff l is given by

$$nDCG\text{-}IA@l = \sum_{i} P(i|q)\ nDCG_i@l \qquad (1)$$

where $nDCG_i@l$ is the "standard" version of nDCG at l[1] computed by assuming that the sole intent of query q is i. That is, for each intent i, we separately imagine an ideal ranked output, which lists up documents that are relevant to intent i in decreasing order of relevance, and compare the system output against the i-th ideal ranked output.

The IA metrics assume that every user has a single intent (i.e. category), and aim to satisfy the "average" user. Suppose that a query has two intents i_1 and i_2 with $P(i_1|q) = 0.9$ and $P(i_2|q) = 0.1$. Then, whether a document relevant to intent i_2 is retrieved or not has very little impact on the overall nDCG-IA value. Hence, a Web search engine tuned with nDCG-IA may fail to include a document relevant to i_2 (say) on the first page. In other words, the IA metrics tend to ignore long tail users. In contrast, although our proposed metrics also utilise $P(i|q)$, they pay attention even to minor intents, as we shall demonstrate in Section 4 with TREC Diversity Task runs.

Another problem with nDCG-IA is that its value is always less than one (unless a single ranked list is ideal for all intents – which is highly unlikely). Hence, comparing and averaging IA metric values across topics are not recommended.

The Diversity Task of the TREC 2009 Web Track used the IA version of precision (precision-IA) as a secondary metric. In addition to the general problems of IA metrics discussed above, precision-IA has a few more weaknesses: (i) its per-intent precision itself is undernormalised, when the document cutoff (e.g. $l = 10$) is larger than the number of relevant documents for the intent[2]; (ii) it disregards the positions of relevant documents. Hence we prefer to use nDCG-IA as one of our "baselines."

2.3 α-nDCG

Clarke *et al.* [6] proposed *α-nDCG* to evaluate diversity and novelty in search results. They view information needs and documents as sets of *nuggets*, based on the ideas from question answering and summarisation evaluation. For example, suppose a query has two intents (or "information needs") i_1 and i_2. Then they represent them in the form $i_1, i_2 \subseteq \{n_1, \dots, n_m\}$ where n_k represents a nugget and m is the total number of nuggets involved for this particular topic. For example, it may be that $m = 4$, $i_1 = \{n_1\}$ and $i_2 = \{n_2, n_3, n_4\}$. We shall come back to this example later.

Let $J_i(k)$ be a flag indicating whether the document at rank k is relevant to intent i or not, and let $C_i(r) = \sum_{k=1}^{r} J_i(k)$, i.e. the number of relevant documents found within top r for intent i. Moreover, for convenience let $C_i(0) = 0$. For each document at rank r, Clarke *et al.* defined what we call *novelty-biased gain $NG(r)$* as follows:

$$NG(r) = \sum_{i=1}^{m} J_i(r)(1-\alpha)^{C_i(r-1)} \qquad (2)$$

where α is a parameter, which we shall discuss later.

Clarke *et al.* [6] computed nDCG based on $NG(r)$, instead of the traditional gain that directly reflects the graded relevance value of a document [8]. The resultant metric is called α-nDCG. When $\alpha = 0$, α-nDCG is reduced to standard nDCG with the number of matching nuggets as the graded relevance value. The key feature of α-nDCG is that it discounts the value of each retrieved relevant document for intent i based on the number of relevant documents already seen for the same intent.

Although α-nDCG is a theoretically-derived metric, the theory builds on a series of assumptions. Firstly, α-nDCG assumes that the relevance of one nugget to the user's intent is independent of other nuggets, that a nugget can either be relevant or nonrelevant to the need, and therefore that only the number of matching nuggets determines the importance of an intent. Hence, unlike the IA metrics, it does not accomodate $P(i|q)$. Secondly, it assumes that the relevance of one nugget to a document is independent of other nuggets, that a nugget can either be relevant or non-

[1] For every rank r, the gain at rank r is discounted by dividing it by $\log(r + 1)$. Note that this differs from the original discounting method that used $\log_b(r)$ with log base b [8].

[2] Of the 199 intents from the TREC 2009 Web Track Diversity Task data (See Table 1), as many as 95 intents have fewer than 10 relevant documents. With $l = 10$, it is impossible to achieve perfect precision for these intents.

relevant to a document, and therefore that only the number of matching nuggets determines the importance of a document. Consider the aforementioned example with i_1 and i_2, having one nugget and three nuggets, respectively. According to α-nDCG, a document that covers both n_2 and n_3 is more important than one that covers n_1, even though the former only partially covers intent i_2 while the latter completely covers i_1. Moreover, consider an ambiguous query such as "java" [20] run against a Wikipedia document collection, and suppose that each of its intents (or *senses*) i_j corresponds to a single nugget n_j. According to α-nDCG, a single disambiguation page that lists up m different senses of "java" retrieved at rank 1 is more valuable than a set of m pages that actually describe each sense in detail, retrieved at ranks 1-m. Thirdly, α-nDCG assumes that relevance assessor's positive judgments (i.e. judging that a document contains a nugget) are erroneous with probability $1 - \alpha$, while negative judgments (i.e. judging that a document does not contain a nugget) are always correct (See Section 4.1 in [6]). This flat probability α is exactly the parameter used for discounting "redundant" documents, as shown in Eq. 2. However, we are inclined to believe that erroneous negative judgments (i.e. missing an existing nugget in a document) are at least as likely as erroneous positive judgments (i.e. reporting a nonexistent nugget in a document).

The fact that α-nDCG has two distinct discount mechanisms also deserves discussions. α-nDCG first discounts a relevant document based on the number of *relevant* documents seen so far for the same intent, using α. This is meant to penalise redundancy (i.e. lack of novelty) in the ranked list. The metric then discounts the same relevant document based on the number of *total* documents seen so far (i.e. the absolute document rank), using the logarithmic discount. It is not clear whether this explicit penalty on redundancy through "double discount" is necessary at least for Web search, where returning a *minimal* set of documents that together cover all intents is not an absolute requirement: we can present (say) 10 documents regardless of whether there are (say) two intents or five intents in a query. Moreover, even if two retrieved documents are relevant to the same intent, the second relevant document may still be informative to the user, unless the two documents are (near-) duplicates. We also point out that one of the motivations for introducing the original logarithmic discount in nDCG was that the user may have "cumulated information from documents already seen" [8]. If this is the case, α-nDCG is discounting "redundant" documents twice, in different ways.

Another potential weakness of α-nDCG is that computing its ideal gain vector is NP-complete, and therefore an approximation is required. Hence, in theory, a system output may outperform the approximated suboptimal ideal output. We believe that evaluation metrics should be easy to understand and easy to compute.

As α-nDCG was the primary metric used in the recent Diversity Task of the TREC 2009 Web Track, we use it as our primary "baseline" in our experiments.

2.4 NRBP

Very recently, Clarke, Kolla and Vechtomova [7] have proposed *Novelty- and Rank-Biased Precision* (NRBP), by combining the ideas of α-nDCG and *Rank-Biased Precision* (RBP) [9]. The most basic version of NRBP is defined as:

$$NRBP = \frac{1-(1-\alpha)p}{m} \sum_{r=1}^{\infty} p^{r-1} \sum_{i=1}^{m} J_i(r)(1-\alpha)^{C_i(r)} \quad (3)$$

where p is the *persistence parameter* of RBP, i.e. the probability that the user will move from one document to the one beneath it, irrespective of document relevance. Hence, just like α-nDCG, NRBP has two discount mechanisms: one from the viewpoint of redundancy based on α, and another from the viewpoint of going down the ranks based on p.

Just like α-nDCG, NRBP relies on the novelty-biased gain (Eq. 2), and therefore all of the above potential weaknesses of α-nDCG also apply to NRBP. In addition, it inherits a potential weakness of RBP: RBP is heavily undernormalised for topics with few relevant documents and does not average well across topics [16][3]. Similarly, NRBP relies on an "ideal ideal vector" [7], where it is assumed that there are infinite number of highly relevant documents. This implies that the maximum NRBP value any system can achieve for a topic with few relevant document is well below 1. Averaging such undernomalised values across topics can be problematic.

Clarke, Kolla and Vechtomova [7] further discuss the integration of the above form of NRBP with the IA approach of Agrawal *et al.* [1]. They envision a topic with mutually exclusive *categories*, and a set of *subtopics* for each category. They propose to use the probabilities $P(i|q)$ at the category level, and use the above form of NRBP at the subtopic level. While such an evaluation protocol is novel and interesting, it remains to be seen whether it is feasible and to what extent can that somewhat complex metric reflect user's perception of the search quality.

2.5 Other Related Metrics

Chapelle *et al.* [3] briefly discussed an extension of their *Expected Reciprocal Rank* (ERR) metric for handling diversity, following the IA approach [1]. Hence we argue that such a metric inherits the potential disadvantages of the IA metrics discussed in Section 2.2. As Chapelle *et al.* note, ERR is a graded-relevance extension of Reciprocal Rank (RR), and a special case of the *Normalised Cumulative Utility* (NCU) [17], discussed below.

Sakai and Robertson [17] examined a family of metrics called NCU, based on the probability of the user abandoning a ranked list at each document rank and the utility of the ranked list given that rank. When it is assumed that the "stopping" probability is uniform across all relevant documents, NCU is reduced to existing simpler metrics such as *Average Precision* (AP), or its graded-relevance extension *Q-measure* which we will use later to design our proposed metrics. In addition to this flat probability distribution, Sakai and Robertson also examined a *rank-biased* probability distribution, based on the assumptions that users abandon the ranked list at a relevant document, and that they are more likely to abandon it near the top ranks. The rank-biased probability distribution is defined based on the number of relevant documents seen so far and resembles the novelty-based discount of α-nDCG, although they did not discuss diversity/novelty. They showed that the flat probability distribution (i.e. emphasising long-tail users who dig deep

[3] For example, in a binary relevance environment with $p = 0.95$, the best-possible RBP for a query with 5 known relevant documents is only .226 [16].

down the ranked list) leads to more stable experimental results than the rank-biased one, even if the latter is closer to the reality.

In his first proposal of Q-measure, the aforementioned graded-relevance IR metric, Sakai [13] applied it to factoid question answering evaluation, given correct answer strings that form several equivalence classes. He proposed that, within a system output (i.e. ranked list of answer strings), only one answer string from each equivalance class should be counted as relevant. This is similar to the idea behind the parameter α of α-nDCG. However, unlike α-nDCG, Sakai's Q-measure with answer equivalence classes used graded relevance for each answer string.

Reciprocal Rank is another metric that is somewhat related to diversity evaluation, as it does not reward retrieval of multiple relevant documents. So are its graded-relevance variants [15], and "1-call at l", which requires systems to return at least one relevant document within top l [4].

3. PROPOSED METRICS

Our goals are (a) to retrieve documents that cover as many intents as possible; and (b) to rank documents that are highly relevant to more popular intents higher than those that are marginally relevant to less popular intents. In Sections 3.1 and 3.2 we propose how to evaluate Properties (a) and (b), respectively, and in Section 3.3 we combine these evaluation metrics.

3.1 I-recall

Our first "proposal" is to use S-recall (see Section 2.1) to evaluate Property (a). We prefer to call it *I-recall* (intent recall) because we are interested in handling queries with multiple possible intents or interpretations. The intents may well be subtopics of an underspecified query, or they may correspond to different *senses* of an ambiguous query. Thus, let I_q denote the complete set of intents for query q. and let $n = |I_q|$. A document may be relevant to n' ($0 \leq n' \leq n$) intents. Let d_r denote the document at rank r, and let $I(d_r)$ denote the set of intents to which document d_r is relevant. For a document cutoff l, define:

$$I\text{-}rec@l = \frac{|\bigcup_{r=1}^{l} I(d_r)|}{n} \tag{4}$$

which is just the proportion of intents covered by the top l documents. Note that if $l < n$, an I-recall of 1 may not be achievable even if some retrieved relevant documents cover multiple intents. In the experiment we report later, the maximum number of intents per topic in our data is $n = 6$, while we use $l = 10$ to follow the Diversity Task at TREC 2009 Web Track. So I-recall ranges between 0 and 1 for all topics.

Suppose that $n = l = 10$, and that System x has ten documents at ranks 1-10, each relevant to exactly one new intent, while System y has exactly one relevant document at rank 1, which covers all ten intents. Note that, in contrast to α-nDCG, I-recall does *not* rate y higher than x. Returning a *minimal* number of documents that together cover all intents [2] is *not* of our concern, and no NP-hard problem is involved here. This is because Web search can return (say) ten documents within the first search result page, regardless of whether it is possible to cover all the intents with just one document. Moreover, as was discussed in 2.3, a single document that covers multiple intents is not necessarily better than a set of documents that each highly satisfies a single intent, at least for our purpose.

I-recall is a binary-relevance metric – for each intent, a document can either be relevant or nonrelevant – and it assumes that each intent is equally important. The idea is to satisfy every user (each corresponding to a different set of intents) at least to some extent within top l. Below, we discuss simple graded-relevance metrics that can reflect the importance of each intent and examine the positions of relevant documents, to complement I-recall.

3.2 div-nDCG and div-Q

Consider a query q with its set of possible intents I_q, and a document d. Let rel be a random binary variable, and define $rel = 1$ for (q, d) iff $\exists i \in I_q$ s.t. $rel = 1$ for (i, d). That is, we say that d is relevant to q if d is relevant to at least one of q's intents, and otherwise it is not relevant. According to the *probability ranking principle* [12, 20], systems should rank documents by $P(rel = 1|q, d)$.

Like the IA metrics, our presumption is that $P(i|q)$ values are available. Moreover, just like the IA metrics, we assume that, for any pair of intents $i, i' \in I_q$, i and i' are *exclusive*: that is, a user searching on q has only one of the possible intents. Under this assumption, we obtain:

$$P(rel = 1|q, d) = \sum_{i \in I_q} P(i|q) P(rel = 1|i, d) \ . \tag{5}$$

We note that this is exactly what Spärck-Jones, Robertson and Sanderson [20] have casually discussed, though not in the context of how to evaluate systems.

We estimate $P(rel = 1|i, d)$ in Eq. 5 based on manual relevance assessments as follows. Suppose, for example, we have four relevance levels so that highly relevant, relevant, partially relevant and judged nonrelevant documents are manually obtained for each intent i (rather than for each query q). Then we may assign a *gain value* [8], e.g. 3,2,1,0, to each type of the above judged documents, respectively. Let $g_i(d)$ denote the gain value of a document d with respect to intent i. Now, let us further assume that $P(rel = 1|i, d) \propto g_i(d)$. That is, we interpret the gain values as statistics that directly reflect the probability of (binary) relevance of a document to an intent[4]. Then from Eq. 5, the probability ranking principle reduces to ranking documents by

$$\sum_{i \in I_q} P(i|q)\ g_i(d) \tag{6}$$

which we call the *global gain* (GG) of document d given query q.

Let $GG(r)$ denote the global gain of the document at rank r, and let the cumulative GG be $CGG(r) = \sum_{k=1}^{r} GG(k)$. Moreover, let $GG^*(r)$ and $CGG^*(r)$ denote the GG and the CGG at rank r for an ideal ranked output (i.e., one that exhaustively lists up the relevant documents in decreasing order of the global gain as defined in Eq. 6). Note that, unlike nDCG-IA that require $n = |I_q|$ distinct ideal ranked lists, we require a single ideal list.

Based on global gains, existing graded-relevance metrics

[4] This is one way to interpret graded relevance assessments. Note that the probability ranking principle is still based on binary relevance.

such as nDCG and Q-measure [17] can be computed[5]:

$$div\text{-}nDCG@l = \frac{\sum_{r=1}^{l} GG(r)/\log(r+1)}{\sum_{r=1}^{l} GG^{*}(r)/\log(r+1)} \quad (7)$$

$$divBR(r) = \frac{\sum_{k=1}^{r} J(k) + \beta \sum_{k=1}^{r} GG(k)}{r + \beta \sum_{k=1}^{r} GG^{*}(k)} \quad (8)$$

$$div\text{-}Q = \frac{1}{R}\sum_{r=1}^{L} J(r) divBR(r) \quad (9)$$

where l is a document cutoff; $J(k)$ is a flag indicating whether the document at rank k is (at least partially) relevant to at least one intent; β is the persistence parameter for the *blended ratio* [17], which combines precision ($\sum_{k=1}^{r} J(k)/r$) and normalised (global) cumulative gain ($\sum_{k=1}^{r} GG(k)/\sum_{k=1}^{r} GG^{*}(k)$);[6] R is the number of documents that are (at least partially) relevant to at least one intent; and L is the size of the system output.

The only difference between traditional nDCG/Q-measure and div-nDCG/Q is that the latter metrics use the global gains (Eq. 6) instead of the raw gains. Just like α-nDCG, div-nDCG and div-Q assume that documents that cover many intents are important. However, unlike α-nDCG, the importance is weighted according to how important each intent is (i.e. $P(i|q)$), and how relevant each document is to an intent (i.e. $g_i(d)$). These metrics reward systems that satisfy Property (b) discussed earlier.

Note that Q-measure and nDCG rely on the number of relevant documents R (the former directly, and the latter indirectly through the use of an ideal ranked list). Some researchers argue that using R for evaluation no longer makes sense in this Web search era as the true R is difficult/impossible to obtain [23]. However, we argue that interpreting R as the number of *known* relevant documents is one reasonable approach: if relevance assessors have already given us 5 relevant documents for Topic 1, and 50 relevant documents for Topic 2, utilising this information for normalisation is indeed useful. For example, metrics that rely on R such as Q-measure are known to be more *discriminative* than precision-oriented metrics like RBP, in that they are more robust to variance across topics (See Section 4.4) [16].

Unlike metrics like α-nDCG and nDCG-IA, the maximum value of div-nDCG is exacly one regardless of the document cutoff value l. Whereas, the maximum value of div-Q is exactly one *provided that* the system output size L is larger than or equal to R (because if $L < R$, obviously the system cannot list up all relevant documents). In traditional ad hoc tasks at evaluation forums like TREC and NTCIR, usually $L >> R$ holds, so this should not be a problem. However, for Web search evaluation environments in which a small cutoff is often used (e.g. $l = 10$ so that the system output is in effect truncated to size $L = 10$), the above undernomalisation problem may occur. Hence, to avoid this, we use the following cutoff-based version of div-Q in this paper:

$$div\text{-}Q@l = \frac{1}{\min(l,R)}\sum_{r=1}^{l} J(r) divBR(r) \quad (10)$$

3.3 Idiv-nDCG and Idiv-Q

I-recall is a simple, binary-relevance, intent-level metric that rewards wide coverage of different intents in the top ranks. div-nDCG and div-Q are simple, graded-relevance, document-level metrics that reward early retrieval of documents that are highly relevant (or more precisely: "highly likely to be relevant") to major intents. As we want both Properties (a) and (b), we simply combine the metrics as follows:

$$Idiv\text{-}nDCG@l = \gamma I\text{-}rec@l + (1-\gamma) div\text{-}nDCG@l \quad (11)$$

$$Idiv\text{-}Q@l = \gamma I\text{-}rec@l + (1-\gamma) div\text{-}Q@l \quad (12)$$

where γ is a parameter. In this paper, we let $\gamma = 0.5$ by default. We use linear combination because we would like to reward systems even when they satisfy only one of the two properties mentioned above. (A harmonic mean of two metrics, for example, would be zero if either of them is zero.)

The Idiv metrics are rather ad hoc, in that they lack a unified user model[7]. Nevertheless, they offer several strengths:

1. They are easy to interpret. It is easy to imagine the kind of ranked output that the Idiv metrics are designed to pursue – all the possible intents should be listed up in the early ranks, and highly relevant documents for the major intents should be ranked above partially relevant documents for the minor intents.

2. They are easy to compute – unlike α-nDCG, S-recall@l_{min}, S-precision and WS-precision (See Section 2.1), no NP-hard problem is involved.

3. Unlike IA metrics, even minor intents are considered important, especially with a large γ ($0 \leq \gamma \leq 1$).

4. Unlike α-nDCG, it accomodates $P(i|q)$ and graded-relevance of documents to each intent.

5. Unlike α-nDCG and IA metrics, they are guaranteed to range fully between 0 and 1, provided that the document cutoff l is chosen for I-recall so that it is no smaller than the maximum number of intents across the query set (See Section 3.1).

4. EXPERIMENTS

4.1 Data

We have already clarified the advantages of our Idiv metrics over existing metrics. We now demonstrate how they work in practice compared to nDCG-IA and α-nDCG, using the recent Diversity Task (Category A) data from the TREC 2009 Web Track [5]. The statistics of the test collection are summarised in Table 1. As it shows, the topic set includes 243 *subtopics*, but only 199 of them have at least one relevant document. These are treated as our *intents*. As many as 23 intents have only one relevant document, and we point out that this is not good news for nDCG-IA, since its per-intent nDCG values for such intents will rely solely on whether one particular document is in top 10 or not.

[5] "div" obviously stands for "diversity."

[6] A small β (e.g. $\beta = 0.1$) makes the blended ratio resemble precision; a large β (e.g. $\beta = 1000$) makes it resemble normalised cumulative gain. The former implies heavier penalties for relevant documents retrieved later in the ranked list. We let $\beta = 1$ in this paper.

[7] The original Q-measure, as an instance of NCU, does have a user model: There is a population of users abandoning the ranked list at different ranks, and the "stopping probability" is uniform across all relevant documents [17].

Table 1: Statistics of the TREC 2009 Web Track Diversity Task (Category A) test collection.

Documents	ClueWeb09 (approx. one billion Web pages; 25TB of uncompressed data in multiple languages) [5]
Topics	50 topics (12 ambiguous; 38 faceted) with subtopics. 243 subtopics (177 informational; 66 navigational)
Intents	199 intents (i.e. subtopics with at least one relevant document).
	Max. #intents per topic: 6. Max. #intents per document: 5.
Relevance data	Total relevant across 50 topics: 4942. Total relevant across 199 intents: 6501.

Table 2: τ and τ_{ap} rank correlation: Idiv-nDCG@10 run rankings vs. other rankings as baselines.

	$\gamma = 0$	$\gamma = .2$	$\gamma = .5$	$\gamma = 0.8$	$\gamma = 1$ (I-rec@10)
α-nDCG@10	.880/.778	**.913**/.807	.893/.747	.873/.747	.860/.729
nDCG-IA@10	**.960/.931**	**.940/.908**	.867/.774	.807/.709	.793/.692
Idiv-Q@10 (same γ)	.880/.882	**.947/.931**	**.980/.977**	**.993/.990**	**1/1**

Table 3: τ and τ_{ap} rank correlation: Idiv-Q@10 run rankings vs. other rankings as baselines.

	$\gamma = 0$	$\gamma = .2$	$\gamma = .5$	$\gamma = 0.8$	$\gamma = 1$ (I-rec@10)
α-nDCG@10	.773/.670	.860/.742	.887/.733	.867/.736	.860/.729
nDCG-IA@10	.880/.846	**.953/.910**	.873/.775	.800/.700	.793/.692
Idiv-nDCG@10 (same γ)	.880/.885	**.947/.934**	**.980/.977**	**.993/.990**	**1/1**

Table 4: τ and τ_{ap} rank correlation: effect of γ on the Idiv-nDCG@10 (Idiv-Q@10) run ranking, with $\gamma = .5$ as the baseline.

	$\gamma = 0$	$\gamma = .2$	$\gamma = 0.8$	$\gamma = 1$ (I-rec@10)
Idiv-nDCG@10 ($\gamma = .5$)	.867/.774	**.913/.812**	**.940/.916**	**.927**/.898
Idiv-Q@10 ($\gamma = .5$)	.807/.722	.893/.792	**.927/.905**	**.907**/.877

Unfortunately, this test collection is not ideal for our purpose, because it has neither the information on $P(i|q)$ for each topic nor graded-relevance assessments for each intent. We overcome the first problem by using simulated $P(i|q)$ values: for a query with n intents, we assume that the j-th intent has the probability $2^{n-j+1}/\sum_{k=1}^{n} 2^k$. As for the second problem, we chose to use the existing binary relevance assessments "as is" even though this means that the graded-relevance capability of the Idiv metrics cannot be demonstrated. Thus, in our experiment, the global gain of a document (Eq. 6) at rank r reduces to:

$$GG_{binary}(r) = \sum_{i \in I_q} P(i|q) J_i(r) \ . \qquad (13)$$

We plan to construct and use graded-relevance diversity test collections in our future work.

The TREC 2009 Web diversity task received 25 run submissions. Below, we evaluate all of these runs with our proposed metrics as well as α-nDCG and nDCG-IA. As was mentioned in Section 3.1, we use $l = 10$ as the document cutoff for all metrics, because diversifying the first result page is especially important for Web search engines. We used the official α-nDCG values released by the Web track organisers (hence $\alpha = .5$); as for all other metrics, we used a set of scripts that we developed ourselves for diversity evalution in general. The scripts are publicly available at
`http://research.nii.ac.jp/ntcir/tools/ir4qa_eval2.tar.gz`.

4.2 System Ranking Comparisons

We quantify the similarity between two run rankings (based on two different evaluation metrics) by *Kendall's τ rank correlation* and Yilmaz/Aslam/Robertson τ_{ap} [21]. τ is a monotonic function of the probability that a *randomly chosen* pair of ranked items is ordered identically in the two rankings. Hence a swap near the top of a ranked list and that near the bottom of the same list has equal impact. Whereas, τ_{ap} is "top-heavy," in that it is a monotonic function of the probability that a randomly chosen item *and one ranked above it* are ordered identically in the two rankings. While τ is symmetrical, τ_{ap} is not: it treats one ranked list as the ground truth.

Table 2 shows τ and τ_{ap} rank correlations between a run ranking based on Idiv-nDCG (with $\gamma = 0, .2, .5, .8, 1$) and one based on another metric. As Eq. 11 shows, Idiv-nDCG with $\gamma = 0$ equals div-nDCG, and Idiv-nDCG with $\gamma = 1$ equals I-recall. Note also that Idiv-nDCG is compared with Idiv-Q using the same value of γ. Values higher than 0.9 are shown in bold for convenience. For example, the τ between Idiv-nDCG ($\gamma = .5$) and α-nDCG is .893, and the corresponding τ_{ap} value (with α-nDCG taken as the ground truth) is .747. It can be observed that:

- Idiv-nDCG with different γ's rank runs similarly to α-nDCG, but there are differences;
- Idiv-nDCG with different γ's rank runs similarly to nDCG-IA, and the similarity increases as γ becomes smaller (i.e. as Idiv-nDCG moves away from I-recall and towards div-nDCG);
- Idiv-nDCG rank runs similarly to Idiv-Q, especially with a large γ (Note that when $\gamma = 1$, they both equal I-recall).

Table 3 shows a similar table for Idiv-Q. It can be observed that Idiv-Q with different γ's also rank runs similarly to α-nDCG and to nDCG-IA, but that there are differences.

Table 4 shows τ and τ_{ap} rank correlations between two rankings, both based on Idiv-nDCG but with different γ's. The table also contains similar information for Idiv-Q. It can be observed that different γ's do not change the run ranking dramatically.

Figure 1 also shows the robustness of Idiv-nDCG to the choice of γ. The two axes represent I-recall and div-nDCG, and the solid slant lines represent contour lines for Idiv-nDCG with $\gamma = .5$; the dotted slant lines represent those with $\gamma = .2$ and $\gamma = .8$. Each circle in the graph represents a run. For example, it can be observed that runs MSDiv3 and MSRACS lie close to the same solid contour line, and therefore that they are almost equally effective in terms of Idiv-nDCG with $\gamma = .5$. MSDiv2 is the top performer according to Idiv-nDCG with $\gamma = .5$ and with $\gamma = .8$; MSDiv3 is the top performer according to Idiv-nDCG with $\gamma = .2$ (which is in agreement with α-nDCG as we shall see later). However, the "top five" runs in Figure 1 lie very close to each

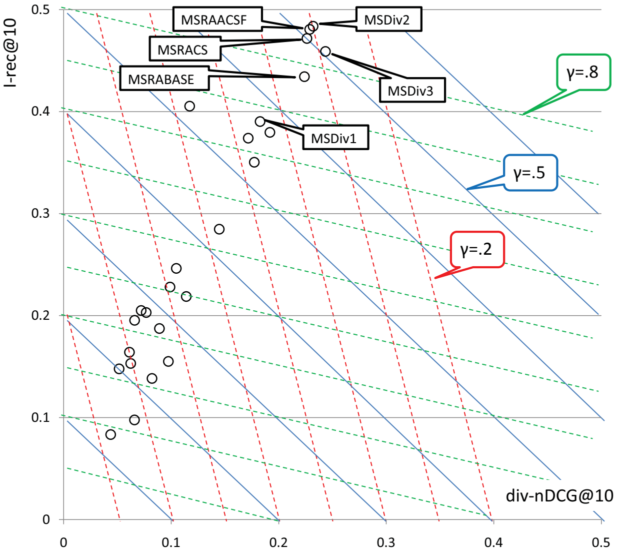


Figure 1: TREC 2009 diversity runs evaluated with Idiv-nDCG@10, with contour lines for $\gamma = .2, .5, .8$.

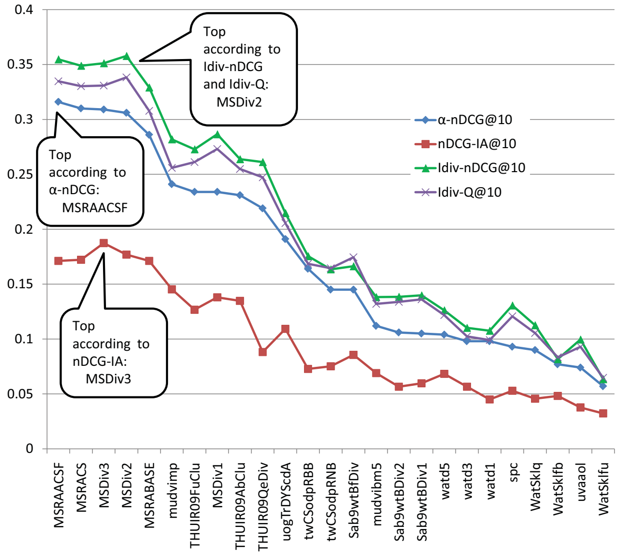


Figure 2: TREC 2009 diversity runs evaluated with different metrics, sorted by the official α-nDCG values.

other and the differences are not statistically significant (See Section 4.4). The main observations from this figure are:

- I-recall (an intent-based recall metric) and div-nDCG (a document-based precision metric) are highly correlated with each other[8];
- Because of the high correlation between I-recall and div-nDCG, Idiv-nDCG is quite robust to the choice of γ for ranking runs.

We have obtained very similar results for Idiv-Q as well.

Since Idiv-nDCG and Idiv-Q are robust to the choice of γ for ranking runs, and because we do not have a systematic method for determining γ, we focus our attention on $\gamma = .5$ henceforth.

Figure 2 visualises the run rankings by Idiv-nDCG, Idiv-Q (both with $\gamma = .5$), nDCG-IA and α-nDCG. The runs have been sorted by the official α-nDCG performance, so the α-nDCG curve is monotonically decreasing; each increase in the other curves represents a disagreement with α-nDCG. As indicated in the figure, Idiv-nDCG and Idiv-Q rank MSDiv2 at the top; nDCG-IA ranks MSDiv3 at the top; and α-nDCG ranks MSRAACSF at the top. However, none of these runs is statistically significantly better than the others in any of the metrics.

We have demonstrated that, given the intent probabilities $P(i|q)$, Idiv-nDCG and Idiv-Q produce run rankings that are similar to, but different from those based on α-nDCG and nDCG-IA. A more important question is whether these metrics are measuring what we want to measure. Below, we examine the actual ranked lists of MSDiv2 and MSRAACSF to demonstrate how our proposed metrics reward diversified systems.

4.3 A Closer Look at Disagreements between Metrics

Figure 3 shows, for Idiv-nDCG, nDCG-IA and α-nDCG, the performance of MSDiv2 *minus* that of MSRAACSF per topic. Thus, dots above the horizontal axis represent "wins" by MSDiv2, and those below represent wins by MSRAACSF. For 38 topics out of 50, the three metrics agree with one another as to which run is better. Idiv-nDCG disagrees with nDCG-IA for 7 topics, and disagrees with α-nDCG for 6 topics; nDCG-IA disagrees with α-nDCG for 9 topics. The trend is similar for Idiv-Q. Below, we take a close look at three topics highted in Figure 3.

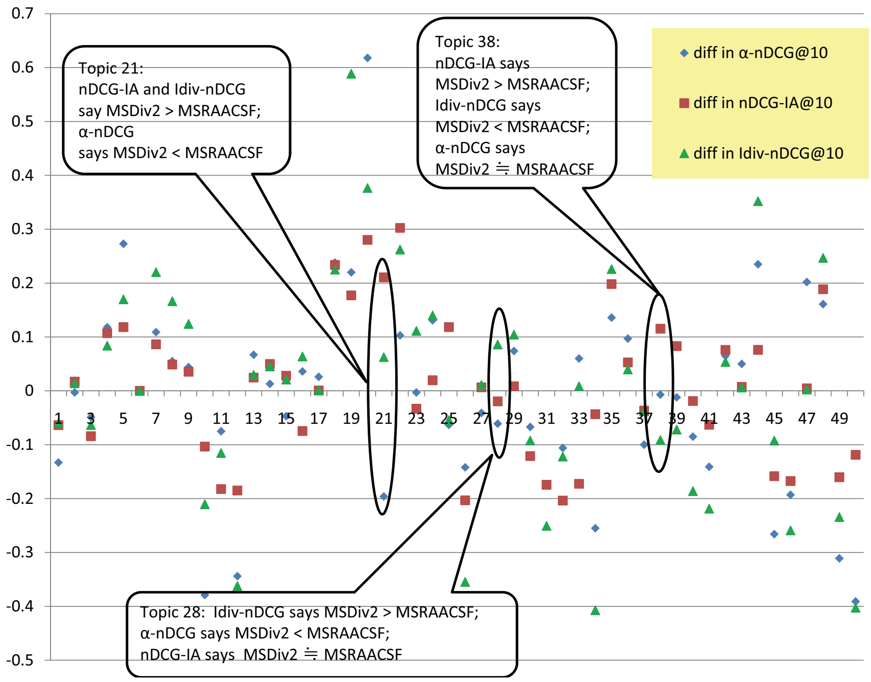
Topic 21 has $n = 5$ intents, with $P(i_1|q) = 32/62$, $P(i_1|q) = 16/62, \ldots, P(i_5|q) = 2/62$. The actual top-10 ranked lists, with relevant intents for each document, are shown below. (Document ID prefixes "clueweb09-en" are omitted throughout this paper to save space.)

rank	MSDiv2		MSRAACSF	
1	0005-48-03496		0010-85-03735	i_5
2	0013-96-08199		0012-66-27529	i_5
3	0013-66-18211		0012-46-18076	
4	0013-66-18147	i_1	0013-27-24714	
5	wp01-79-16252		wp02-04-16080	
6	0003-59-13504		0037-28-21832	
7	0041-68-19894		0062-72-09421	
8	0004-64-17952		0047-54-35373	
9	0080-91-29662		0099-86-05525	
10	0004-51-16201		0060-08-24587	

Because Idiv-nDCG and nDCG-IA utilise the fact that i_1 is a much more popular intent than i_5 (in our experimental setting), they prefer MSDiv2 to MSRAACSF. Whereas, α-nDCG prefers MSRAACSF, as it can only treat each intent equally. This example demonstrates that, given the intent probabilities, Idiv-nDCG and nDCG-IA can be more intuitive than α-nDCG.

Topic 28 also has $n = 5$ intents. The ranked lists are shown below:

[8] $\tau = .793$. $\tau_{ap} = .686$ with I-recall as the ground truth; $\tau_{ap} = .749$ with div-nDCG as the ground truth.



Topic 21: nDCG-IA and Idiv-nDCG say MSDiv2 > MSRAACSF; α-nDCG says MSDiv2 < MSRAACSF

Topic 38: nDCG-IA says MSDiv2 > MSRAACSF; Idiv-nDCG says MSDiv2 < MSRAACSF; α-nDCG says MSDiv2 ≒ MSRAACSF

Topic 28: Idiv-nDCG says MSDiv2 > MSRAACSF; α-nDCG says MSDiv2 < MSRAACSF; nDCG-IA says MSDiv2 ≒ MSRAACSF

Figure 3: Per-topic performance of MSDiv2 minus that of MSRAACSF.

rank	MSDiv2		MSRAACSF	
1	wp00-88-23480		0022-00-34499	i_4
2	0007-19-33611	i_4	0007-19-33611	i_4
3	0002-47-07470		wp00-88-23480	
4	0009-64-18412	i_4	0037-28-16212	i_2
5	wp02-24-00069		0009-64-18412	i_4
6	wp01-05-00069		0029-92-16034	
7	0127-32-28190		0064-52-31354	
8	0004-42-16367		0115-25-00480	
9	0100-95-14715	i_5	0034-67-36999	
10	0030-22-33316	i_2	0020-05-14160	

Idiv-nDCG prefers MSDiv2; α-nDCG prefers MSRAACSF; nDCG-IA is undecided, but leans towards MSRAACSF. Since MSDiv2 covers three intents while MSRAACSF covers only two, Idiv-nDCG may be the most intuitive for the purpose of diversification. Note that having a document relevant to i_5 in MSDiv2 has very little impact on the computation of nDCG-IA as this is a "very minor" intent (See Eq. 1).

Topic 38 has $n = 3$ intents, with $P(i_1|q) = 8/14$, $P(i_2|q) = 4/14$, $P(i_5|q) = 2/14$. The ranked lists are shown below:

rank	MSDiv2		MSRAACSF	
1	0011-83-04353		0022-32-01173	
2	0022-89-07369	i_1	0011-83-04353	
3	0009-04-14265		0009-04-13849	
4	0006-15-25073	i_1	0008-94-08357	
5	0009-49-19901	i_1	0039-80-34409	i_1
6	0078-29-22726		0004-30-16737	i_1, i_3
7	0026-63-30995		0009-70-21188	
8	0052-82-45572	i_1	0024-34-31519	i_1
9	0008-62-07163	i_1	0035-17-17820	
10	0004-07-33887	i_1	0009-12-14653	i_1

Idiv-nDCG prefers MSRAACSF, as it covers not only i_1 but also i_3. α-nDCG is undecided, but leans towards MSRAACSF. Again, nDCG-IA prefers MSDiv2 as it virtually ignores the minor intent i_3.

The above examples demonstrate that:

- nDCG-IA indeed tends to ignore minor intents and may be counterintuitive for the purpose of evaluating diversity;
- Given the intent probabilities, Idiv-nDCG can be more intuitive than α-nDCG, as the latter metric does not take the probabilities into account. (The same is true for Idiv-Q.)

We also remind the reader that our proposed metrics (a) can handle per-intent graded relevance (unlike α-nDCG), and (b) do not require approximation and range fully between 0 and 1 (unlike α-nDCG and nDCG-IA).

4.4 Discriminative Power

We finally compare our proposed metrics with α-nDCG and nDCG-IA in terms of *discriminative power* [14]. We want metrics that are robust to variation across topics, so that the same conclusion can be reached as to which of two given systems is better, regardless of the choice of the topic set. More precisely, we measure discriminative power by conducting a statistical significance test for every pair of runs, and counting the number of significantly different pairs. We have 25 runs, so 25*24/2=300 run pairs are tested. For significance testing, we use the two-tailed *paired bootstrap test*, with $\alpha = .05$ and $B = 1000$ *bootstrap samples* [14]. Note that this experiment is not about whether the metrics are right or wrong; it is about how metrics can be consistent across experiments and as a result how often differences between systems can be detected with high confidence. We regard high discriminative power as a necessary condition for a good evaluation metric, not as a sufficient condition.

The discriminative power method also provides a natural estimate of the performance delta between two systems required to achieve statistical significance. This is done by recording, for every run pair, the performance difference that corresponds to the borderline between significance and non-significance among the $B = 1000$ trials, and then by selecting the largest value among all run pairs (i.e. a conservative estimate). This is one of the advantages of using the boot-

Table 5: Discriminative power and estimated difference between two mean performances required for $\alpha = .05$.

metric	disc. power	diff. required
Idiv-Q@10	192/300=64.0%	0.12
Idiv-nDCG@10	188/300=62.7%	0.13
α-nDCG	185/300=61.7%	0.10
I-rec@10	181/300=60.3%	0.20
nDCG-IA	173/300=57.7%	0.08
div-nDCG@10	171/300=57.0%	0.10
div-Q@10	148/300=49.3%	0.12

strap test. Details can be found in [14].

Table 5 summarises the results of our discriminative power experiments. For example, Idiv-Q is the most discriminative metric of the ones we examined, managing to detest statistical significance for 192 run pairs; Given 50 topics, when the performance difference between to systems is 0.12 or higher in Idiv-Q, this difference is usually statistically significant. It can be observed that[9]:

- Idiv-Q ($\gamma = .5$) is more discriminative than its components, namely I-recall and div-Q; similarly, Idiv-nDCG ($\gamma = .5$) is more discriminative than I-recall and div-nDCG.
- Idiv-Q and Idiv-nDCG ($\gamma = .5$) are at least as discriminative as α-nDCG ($\alpha = .5$) and possibly more discriminative than nDCG-IA. For example, Idiv-Q manages to obtain seven more significantly different run pairs compared to α-nDCG.

Hence Idiv-Q and Idiv-nDCG are at least as good as existing metrics from the viewpoint of robustness to variation across topics and hence consistency across experiments, and our approach of combing I-recall and div-Q/div-nDCG is probably beneficial.

5. CONCLUSIONS

This paper proposed simple, intuitive metrics called Idiv-nDCG and Idiv-Q for evaluating diversified search results. They are designed to (a) retrieve documents that cover as many intents as possible; and (b) rank documents that are highly relevant to more popular intents higher than those that are marginally relevant to less popular intents. We showed that they offer several advantages over existing metrics: They do not require approximation and range fully between 0 and 1; Unlike the IA metrics, they pay attention to minor intents; Unlike α-nDCG, they can accomodate $P(i|q)$ and per-intent graded relevance. We have also demonstrated that they correlate well with the existing metrics, *and* can be more intuitive. We therefore argue that they are suitable for evaluation with diversity test collections that accomodate either $P(i|q)$ or per-intent graded relevance.

Our future work includes setting parameters such as γ based on clickthrough data (possibly per topic); explicitly handling informational vs. navigational intents; and designing new diversity metrics that are directly motivated by a user model. However, we maintain that evaluation metrics should be easy to understand and easy to compute.

[9] According to these bootstrap tests, the "top five" runs in Figure 1 are not significantly different from one another in terms of any of the metrics shown in Table 5. Whereas, some of these runs are significantly better than MSDiv1 also shown in Figure 1 in terms of several different metrics.

6. REFERENCES

[1] R. Agrawal, G. Sreenivas, A. Halverson, and S. Leong. Diversifying search results. In *Proceedings of ACM WSDM 2009*, pages 5–14, 2009.

[2] B. Carterette and P. Chandar. Probabilistic models of novel document rankings for faceted topic retrieval. In *Proceedings of ACM CIKM 2009*, pages 1287–1296, 2009.

[3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of ACM CIKM 2009*, pages 621–630, 2009.

[4] H. Chen and D. R. Karger. Less is more. In *Proceedings of ACM SIGIR 2006*, pages 429–436, 2006.

[5] C. L. Clarke, N. Craswell, and I. Soboroff. Preliminary report on the TREC 2009 web track. 2009.

[6] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of ACM SIGIR 2008*, pages 659–666, 2009.

[7] C. L. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Advances in Information Retrieval Theory (ICTIR 2009), LNCS 5766*, pages 188–199, 2009.

[8] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[9] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):Article No.2, 2008.

[10] P. Over. TREC-7 interactive track report. In *Proceedings of TREC-7*, 1999.

[11] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43(2):46–52, 2009.

[12] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:130–137, 1977.

[13] T. Sakai. New performance metrics based on multigrade relevance: Their application to question answering. In *NTCIR-4 Proceedings Open Submission Session*, 2004.

[14] T. Sakai. Evaluating information retrieval metrics based on bootstrap hypothesis tests. *IPSJ Transactions on Databases*, 48(SIG9 (TOD35)):11–28, 2007.

[15] T. Sakai. On the properties of evaluation metrics for finding one highly relevant document. *IPSJ Transactions on Databases*, 48(SIG9 (TOD35)):29–46, 2007.

[16] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11:447–470, 2008.

[17] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA 2008*, pages 30–41, 2008.

[18] M. Sanderson. Ambiguous queries: Test collections need more sense. In *Proceedings of ACM SIGIR 2008*, pages 499–506, 2008.

[19] R. Song, D. Qi, H. Liu, T. Sakai, J.-Y. Nie, H.-W. Hon, and Y. Yu. Constructing a test collection with multi-intent queries. In *Proceedings of EVIA 2010*, 2010.

[20] K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: Implications for retrieval tests, systems and theories. *SIGIR Forum*, 41(2):8–17, 2007.

[21] E. Yilmaz, J. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of ACM SIGIR 2008*, pages 587–594, 2008.

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

[23] J. Zobel, A. Moffat, and L. A. Park. Against recall: Is it persistence, cardinality, density, coverage, or totality? *SIGIR Forum*, 43(1):3–15, 2009.