Evaluating Diversified Search Results Using Per-intent Graded Relevance

Tetsuya Sakai
Microsoft Research Asia, Beijing, P.R.C.
tetsuyasakai@acm.org

Ruihua Song
Microsoft Research Asia, Beijing, P.R.C.
Song.Ruihua@microsoft.com

ABSTRACT

Search queries are often ambiguous and/or underspecified. To accomodate different user needs, search result diversification has received attention in the past few years. Accordingly, several new metrics for evaluating diversification have been proposed, but their properties are little understood. We compare the properties of existing metrics given the premises that (1) queries may have multiple intents; (2) the likelihood of each intent given a query is available; and (3) graded relevance assessments are available for each intent. We compare a wide range of traditional and diversified IR metrics after adding graded relevance assessments to the TREC 2009 Web track diversity task test collection which originally had binary relevance assessments. Our primary criterion is discriminative power, which represents the reliability of a metric in an experiment. Our results show that diversified IR experiments with a given number of topics can be as reliable as traditional IR experiments with the same number of topics, provided that the right metrics are used. Moreover, we compare the intuitiveness of diversified IR metrics by closely examining the actual ranked lists from TREC. We show that a family of metrics called D♯-measures have several advantages over other metrics such as α-nDCG and Intent-Aware metrics.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

ambiguity, diversity, evaluation, graded relevance, test collection

1. INTRODUCTION

Traditional information retrieval (IR) research has mostly focussed on satisfying clearly specified information needs. However, in Web search, queries are often *ambiguous* and/or *underspecified* [9]. When a search engine has no or little knowledge of the user, the best it can do may be to produce an output that reflects several *interpretations* (or *intents*) of such queries to accomodate a large population of users. In light of this, *search result diversification* is beginning to receive attention [1, 6, 10, 15, 24]. While modern search engines can display multiple blocks of information, some textual and others nontextual, a ranked list of URLs remains the primary output feature today. Hence how to appropriately evaluate a diversified list of URLs is an important research problem for advancing the state-of-the-art of search engines.

A popular method for evaluating *traditional* (i.e. non-diversified) IR is to use *normalised Discounted Cumulative Gain* (nDCG) [11] based on *graded relevance* assessments, taking into account the fact that some relevant documents are more relevant than others. Assuming that we can list up some possible intents for a given ambiguous or underspecified query in advance, a natural extension of the above evaluation framework would be to hire assessors to conduct graded relevance assessments *for each intent* of that query. For example, for the ambiguous query "apple," we should be able to collect documents that are highly/marginally relevant to the intent "Apple the Steve Jobs company," and those that are highly/marginally relevant to the intent "apple the fruit." Moreover, if we know that *some intents are more likely than others*, we should probably reflect that information in evaluating search engines. For example, if a search query log suggests that users are more likely to search for "Apple the company" than for "apple the fruit," the search engine may choose to return more URLs relevant to the former than ones relevant to the latter.

Recently, several evaluation metrics for handling search result diversification have been proposed. Surprisingly, however, not all of them accomodate *per-intent graded relevance* and *intent probabilities*. The objective of this study is to compare the reliability of diversified IR metrics as well as traditional IR metrics on which the diversity metrics are based. By reliability, we informally mean the ability of a metric to detect "real" performance differences as opposed to those observed by chance. As a measure of reliability, we use *discriminative power* [19] which has been used in several evaluation studies over the past five years [7, 17, 20, 25, 29]. In addition, we closely examine some actual ranked lists from TREC and discuss the intuitiveness of different diversified IR metrics.

To our knowledge, our work is the first to use a diversity test collection with per-intent graded relevance data for the purpose of comparing diversified IR metrics. This is somewhat surprising, as many of the existing diversity metrics can handle per-intent graded relevance. We first add graded relevance assessments to the TREC 2009 Web track diversity task test collection which originally had binary relevance assessments. We then examine a wide variety of metrics: six for traditional IR, and eleven for diversified IR. We show that a family of metrics called D♯-measures have several advantages over *α-nDCG* [8] and *Intent-Aware* metrics [1], including

SIGIR'11, July 24–28, 2011, Beijing, China.

high discriminative power and intuitiveness. We also show that the "best" diversified IR metrics can be as discriminative as the "best" traditional IR metrics, given the same number of topics.

The remainder of this paper is organised as follows. Section 2 discusses previous work related to the present study, and Section 3 formally defines the traditional and diversified IR metrics that we examine. Section 4 describes discriminative power, our primary criterion for choosing reliable metrics. Section 5 describes our test collections with per-intent graded relevance data, and Section 6 reports on our experiments. Finally, Section 7 concludes this paper.

2. RELATED WORK

Zhai, Cohen and Lafferty [31] used *subtopic recall* for evaluating *subtopic retrieval*. In the present study, we evaluate subtopic recall, but call it *intent recall*. Zhai, Cohen and Lafferty also defined *S-precision* and *WS-precision* at a given subtopic recall level, but the computation of these two metrics involves an NP-hard problem and an approximation is required. Carterette and Chandar [4] evaluated *faceted topic retrieval*, a task similar to subtopic retrieval. Their evaluation metric was subtopic recall at l_{min}, where l_{min} is the minimum rank at which perfect S-recall can be achieved. But they point out that computing l_{min} is NP-hard.

Clarke *et al.* [8] proposed a metric called α-nDCG for evaluating diversified search results. They view both intents and documents as sets of *nuggets*, and assume that the number of nuggets covered by a document determines the graded relevance of that document. Prior to discounting the *gain* of a document based on its rank (as traditional nDCG does [11]), α-nDCG discounts the gain based on "nuggets already seen." This metric was one of the official metrics used at the TREC 2009 Web track diversity task [6], and the present study uses the official α-nDCG values from TREC for computing its discriminative power.

Clarke, Kolla and Vechtomova [9] proposed *Novelty- and Rank-Biased Precision* (NRBP) by combining the ideas of α-nDCG and *Rank-Biased Precision* [13]. NRBP was used at the recent TREC 2010 Web track diversity task. However, Sakai *et al.* [20] argue that NRBP inherits the weaknesses of both α-nDCG and RBP: one practical weakness of (N)RBP is that it is heavily undernomalised for topics with few relevant documents and does not average well. Moreover, the advantage of NRBP over α-nDCG is not clear (at least in terms of discriminative power) in subsequent experiements reported by Clarke *et al.* [7]. We therefore do not consider NRBP.

Agrawal *et al.* [1] proposed *Intent-Aware* (IA) metrics for evaluating diversified search results. They were the first to explicitly incorporate intent probabilities in IR evaluation. Their approach is simple: compute a traditional metric *for each intent* and then finally take an expectation based on the intent probabilities. Clarke, Kolla and Vechtomova [9] have discussed the possibility of evaluation with IA versions of α-nDCG and NRBP.

Chapelle *et al.* [5] proposed *Expected Reciprocal Rank* (ERR) for traditional IR evaluation, and claimed that its IA version can handle diversified IR evaluation. *ERR-IA* was used at the TREC 2010 Web track diversity task. The essence of ERR is that relevant documents are discounted based on the number of *relevant* documents already seen rather than the absolute document ranks.

Robertson, Kanoulas and Yilmaz [17] recently proposed a metric for traditional IR evaluation called *Graded Average Precision* (GAP). GAP assumes that the user has a binary notion of relevance, but that different users have different thresholds over the relevance levels. The present study examines (for the first time) an IA version of GAP, as well as its normalised version.

Sakai *et al.* [20] proposed an alternative method for evaluating diversified search results. Their key idea is to define (global) graded relevance by combining intent probabilities and per-intent graded relevance. Based on an ideal ranked list thus defined, they introduced a family of metrics which we call *D-measures* and *D♯-measures*. Using the TREC 2009 Web diversity task data which have per-intent *binary* relevance assessments, they reported that D♯-measures perform at least as well as α-nDCG and intent recall in terms of discriminative power, while solving some shortcomings of α-nDCG and IA metrics.

Using the same TREC 2009 data set, Clarke *et al.* [7] also compared different diversified IR metrics (with and without "collection-dependent" normalisation) in terms of discriminative power. They pointed out that IA metrics do not necessarily reward high intent recall. Surprisingly, all of the diversified IR metrics they examined (including α-nDCG and NRBP) underperformed the simple intent recall in terms of discriminative power.

The present study is similar to the work by Sakai *et al.* [20] and Clarke *et al.* [7], but has the following new contributions: (a) It is the first to compare different diversified IR metrics *when per-intent graded relevance assessments are available*. (b) It examines the most extensive set of diversified IR metrics, including (normalised) GAP-IA which is being examined for the first time. All of these metrics can handle graded relevance: hence, in previous work, the full potential of these metrics for diversity evaluation has not been demonstrated. (c) While the previous two studies considered either a uniform [7] or non-uniform [20] intent probability distribution, this study considers both. (d) This study investigates the effect of measurement depth on discriminative power, in contrast to previous work which only considered a document cutoff of 10 [20] or 20 [7]. (e) It uses multiple test collections for evaluating *traditional* IR metrics on which our diversified IR metrics are based.

By employing Mechanical Turk users, Sanderson *et al.* [23] examined the *predictive power* of metrics: if a metric prefers one ranked list over another, does the user also prefer the same list? Using the same TREC 2009 Web diversity task data, they examined some traditional and diversified IR metrics. One of their findings was that diversified IR metrics agree reasonably well with human preferences. Clearly, discriminative power is not the only way to evaluate evaluation metrics (See Section 4), and other approaches, especially those that rely on human subjects (e.g. predictive power), should complement our work.

More recently, Brandt *et al.* [3] proposed a framework for presenting a tree of retrieved URLs dynamically, which includes an evaluation method that is a tree version of the IA approach. This probably deserves an investigation in terms of the user's physical and cognitive load when compared to our flat-list approach.

3. EVALUATION METRICS

3.1 Basic Metrics

First, we formally define some graded-relevance evaluation metrics that have been designed for *traditional* IR evaluation.

Our first premise is that we have relevance levels $\{0, \ldots, h\}$, with 0 representing nonrelevance and h representing the highest level. Hence $h = 1$ implies a binary relevance environment. We say that a document is Lx-relevant if its relevance level is x ($0 < x \leq h$). Let R_x denote the number of Lx-relevant documents for a topic and let $R = \sum_x R_x$. Let $J(r) = 1$ if a document at rank r is Lx-relevant ($x > 0$) and $J(r) = 0$ otherwise. Let $C(r) = \sum_{k=1}^{r} J(k)$.

Let GV_x denote the *gain value* for retrieving an Lx-relevant document [11]. Let $g(r) = GV_x$ if a document at rank r is Lx-relevant and $g(r) = 0$ otherwise. Further, let $cg(r) = \sum_{k=1}^{r} g(k)$. We call $g(r)$ and $cg(r)$ the (cumulative) gain at rank r. Also, let

$g^*(r)$ and $cg^*(r)$ denote the (cumulative) gain at rank r in an *ideal ranked list*, obtained by exhaustively listing up Lx-relevant documents in descending order of relevance levels.

We define nDCG and *Q-measure* at document cutoff l as follows [20]:

$$nDCG@l = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \quad (1)$$

$$Q@l = \frac{1}{\min(l,R)} \sum_{r=1}^{l} J(r) \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} \quad (2)$$

where β (≥ 0) is a user persistence parameter for Q. ($\beta = 0$ reduces Q to binary Average Precision.) We let $\beta = 1$ throughout this paper.

Next, we discuss Graded Average Precision (GAP) [17]. GAP assumes that the user considers only relevance levels $x, \ldots, h$ as relevant ($x > 0$) and the rest as nonrelevant with probability p_x, and that $\sum_{x=1}^{h} p_x = 1$. Following a recommendation by Robertson, Kanoulas and Yilmaz [17], we consider a uniform probability distribution: $p_x = 1/h$ for $x \geq 1$. In this particular case, GAP can be defined as follows.

Let $X(r) \in \{0, \ldots, h\}$ denote the relevance level of a document at rank r, and let $M(r,k) = \min(X(r), X(k))$ for any pair of ranks (r,k). Then the *Expected Precision* (EP) [17] at rank r is given by $EP(r) = \frac{1}{r} \sum_{k=1}^{r} \sum_{x=1}^{M(r,k)} p_x$. Under the uniformity assumption, it can be rewritten as:

$$EP(r) = \frac{1}{h\,r} \sum_{k=1}^{r} \sum_{x=1}^{M(r,k)} = \frac{1}{2h\,r} \sum_{k=1}^{r} M(r,k)(M(r,k)+1) \,. \quad (3)$$

Similarly, under the same assumption, the maximum possible value of cumulated EP is: $\sum_{x=1}^{h} R_x \sum_{y=1}^{x} p_y = \frac{1}{2h} \sum_{x=1}^{h} R_x \; x(x+1)$. Therefore, GAP can be expressed as:

$$GAP = \frac{\sum_{r=1}^{\infty} EP(r)}{\frac{1}{2h} \sum_{x=1}^{h} x(x+1) R_x} \quad (4)$$

$$= \frac{\sum_{r=1}^{\infty} \frac{1}{r} \sum_{k=1}^{r} M(r,k)(M(r,k)+1)}{\sum_{x=1}^{h} R_x \; x(x+1)} \,. \quad (5)$$

Note that the above denominator represents a summation over all Lx-relevant ($x > 0$) documents, and therefore can also be written as $\sum_{r=1}^{R} X^*(r)(X^*(r)+1)$, where $X^*(r)$ is the relevance level of a document at rank r in an ideal output. For the purpose of Web search evaluation where the number of documents to be evaluated is typically very small (e.g. $l = 10$), a division by the above sum over *all* relevant documents yields a heavily undernormalised metric. (It is usually impossible to list up all relevant documents if there is space for only 10 URLs.) Hence, we also define *normalised GAP* (nGAP) for evaluation with a small document cutoff:

$$nGAP@l = \frac{\sum_{r=1}^{l} \frac{1}{r} \sum_{k=1}^{r} M(r,k)(M(r,k)+1)}{\sum_{r=1}^{l} X^*(r)(X^*(r)+1)} \,. \quad (6)$$

Next, we define *Expected Reciprocal Rank* (ERR) [5]. Let $Pr(r)$ denote the relevance probability of a document at rank r. Let $dsat(r) = \prod_{k=1}^{r}(1 - Pr(k))$, which is interpreted as the probability that the user is dissatisfied with documents from ranks 1 to r. Then ERR is defined based on the expected probability that the user is finally satisfied at rank r:

$$ERR = \sum_{r=1}^{\infty} \frac{Pr(r)\,dsat(r-1)}{r} \,. \quad (7)$$

Following Chapelle *et al.* [5], we let $Pr(r) = (2^{X(r)} - 1)/2^h$. Note that this makes ERR a very top-heavy metric: any ranked list that has a document of the highest relevance level (h) at rank 1 receives an ERR of $(2^h - 1)/2^h$ or higher. Thus, if we have $h = 3$ relevance levels, the ERR is $7/8 = .875$ or higher; if $h = 4$, the ERR is .938 or higher. Also, in accordance with the above setting for ERR, we let $g(r) = 2^{X(r)} - 1$ for nDCG and Q. This gain value setting appears to be the *de facto* standard for nDCG[1].

Finally, for completeness, we also consider *normalised ERR* [7] for evaluation with a small document cutoff:

$$nERR@l = \frac{\sum_{r=1}^{l} Pr(r)\,dsat(r-1)/r}{\sum_{r=1}^{l} Pr^*(r)\,dsat^*(r-1)/r} \quad (8)$$

where $Pr^*(r)$ is the relevance probability of a document at rank r in an ideal ranked list and $dsat^*(r) = \prod_{k=1}^{r}(1 - Pr^*(k))$. However, we expect the impact of normalisation for ERR to be small since unnormalised ERR already tends to take very large values, as we have discussed earlier.

3.2 α-nDCG

Clarke *et al.* [8] proposed α-nDCG for evaluating diversified search results. They view information needs (i.e. intents) and documents as sets of *nuggets*. In their framework, the assessor judges whether each document contains a nugget or not (i.e. makes a binary decision). Let $J_n(r) = 1$ if a document at rank r is relevant to the n-th nugget and 0 otherwise; let $C_n(r) = \sum_{k=1}^{r} J_n(r)$, i.e. the number of documents observed within top r that contained the n-th nugget. Then *novelty-biased gain* $NG(r)$ is defined as:

$$NG(r) = \sum_{n=1}^{m} J_n(r)(1-\alpha)^{C_n(r-1)} \quad (9)$$

where m is the total number of nuggets for the query and $\alpha (< 1)$ is a parameter. Then α-nDCG can be defined by replacing the raw gain values $g(r)$ and $g^*(r)$ in Eq. 1 with the novelty-biased gains. Thus, α-nDCG discounts gains first based on "nuggets already seen," and then based on document ranks. Note that Eq. 9 defines graded relevance simply based on the number of nuggets that a document covers as well as nugget novelty.

Computing the ideal ranked list for α-nDCG is NP-complete [8] and an approximation is required. In this study, we simply use the official α-nDCG values reported at the TREC 2009 Web diversity task where $\alpha = .5$ [6]. Note that this setting of α corresponds to the assumption that the assessor "finds" a nonexistent nugget in a document 50% of the time but never misses an existing nugget, which is arguably counterintuitive [20].

3.3 Intent-Aware (IA) Metrics

Agrawal *et al.* [1] proposed a simple methodology for evaluating diversified search results. First, we assume that, given a query q with several different intents i, the probability of each intent $Pr(i|q)$ can be estimated, where $\sum_i Pr(i|q) = 1$. Second, we assume that document relevance assessments are available for each intent. Then, for example, nDCG for a particular intent i (nDCG$_i$) can be computed first, and finally *nDCG-IA* can be computed as:

$$nDCG\text{-}IA@l = \sum_i Pr(i|q)\, nDCG_i@l \,. \quad (10)$$

[1]GAP relies on a different notion of graded relevance and its graded relevance setting in our paper is not strictly equivalent to that we use for nDCG, Q and ERR. In the case of GAP with $h = 3$, for example, the uniform probability distribution implies that all users regard $L3$-relevant as relevant, while 67% of them regard $L2$-relevant as relevant, and only 33% regard $L1$-relevant as relevant [17].

Thus, unlike α-nDCG, IA metrics accomodate both intent probabilities and per-intent graded relevance[2]. Other IA metrics can be computed similarly: in the present study, we consider nDCG-IA, (n)GAP-IA and (n)ERR-IA.

One of the disadvantages of IA metrics is that their maximum value is not 1: a single ranked list is almost never ideal for every intent [20]. More importantly, IA metrics tend to be counterintuitive, in that they practically disregard minor intents (i.e. those with relatively low $Pr(i|q)$ values) [7, 20].

3.4 D-measures and D♯-measures

Sakai *et al.* [20] proposed an alternative way to evaluate diversified search results, given intent probabilities and per-intent graded relevance assessments. This solves the undernormalisation problem of IA metrics and also includes a mechanism for explictly boosting *intent recall*, i.e. number of intents covered by a ranked list.

Let us say that document d is relevant to q iff it is relevant to at least one of its intents ($\in \{i\}$). If we assume that the intents are *exclusive* (i.e. a user searching with q has exactly one intent), then the *Probability Ranking Principle* (PRP) [16] reduces to ranking documents by $\sum_i Pr(i|q)Pr(rel = 1|i, d)$, where $Pr(rel = 1|i, d)$ is the probability that document d is relevant to i. Note that relevance is a binary notion here. Let $GV_{i,d}$ denote a gain value for document d with respect to intent i. If we can assign these values so that $Pr(rel = 1|i, d) \propto GV_{i,d}$, then the PRP reduces to ranking documents by $\sum_i Pr(i|q)GV_{i,d}$. This defines a *globally ideal* ranked list, which retrieves documents that are highly relevant to major intents above those that are marginally relevant to minor intents. Based on this ideal list, cumulative-gain-based metrics such as nDCG and Q can be computed, by replacing the raw gain $g(r)$ discussed in Section 3.1 with the *global gain*:

$$GG(r) = \sum_i Pr(i|q)g_i(r) \quad (11)$$

where $g_i(r)$ is the gain value for document at rank r for intent i. Following Sakai *et al.*, we apply this idea to nDCG and Q, and call the resultant metrics *D-nDCG* and *D-Q* (D stands for diversity). Moreover, we collectively call such metrics *D-measures*.

Note that D-measures rely on a single ideal ranked list (in contrast to IA metrics which use multiple "locally ideal" lists), and that the relevance levels are defined more dynamically and implicitly than for traditional nDCG and Q. For example, suppose that we have three *local* (i.e. per-intent) relevance levels $L1$-$L3$. Then, for a topic with three intents, we will have at most $3*3 = 9$ *global* relevance levels. The number of global relevance levels will differ for other topics. Thus, an ideal list is defined not based on a discrete set of relevance levels (which is usually pre-defined over the entire test collection), but based on a sort by the global gain, defined per topic.

Apart from the fact that D-measures avoid the undernomalisation problem of IA metrics by relying on a single "globally ideal" list, these two metric families are quite similar. However, Sakai *et al.* [20] also proposed a simple method to explicitly encourage high intent recall in a search output within the D-measure framework. Let $I\text{-}rec@l$ denote the intent recall at document cutoff l. Then *D♯-measure* ("dee sharp measure") is defined as[3]

$$D\sharp\text{-}measure@l = \gamma I\text{-}rec@l + (1-\gamma)D\text{-}measure@l \quad (12)$$

where γ is a parameter. As Sakai *et al.* showed that the effect of the choice of γ on IR experiments is relatively small due to the fact that I-rec and D-measures are already highly correlated with each other (evidence will be given in Section 6.2), we let $\gamma = .5$ throughout this study. We examine *D♯-nDCG* and *D♯-Q* in this paper. Provided that the document cutoff l is larger or equal to the maximum number of intents, I-rec ranges between 0 and 1, and therefore D♯-measures also range between 0 and 1.

4. DISCRIMINATIVE POWER

Our primary method of comparing evaluation metrics is *discriminative power* [19]. 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 different pairs of runs, and counting the number of significantly different pairs. In this study, we randomly sample 20 runs from each test collection so 20*19/2=190 run pairs are tested in each case[4]. For significance testing, we use the two-tailed *paired bootstrap test*, with 1,000 bootstrap samples [19]. 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. Later in this paper, we shall complement our discriminative power results by examining the actual ranked lists and comparing the intuitiveness of different diversity IR metrics.

It has been pointed out that discriminative power is not useful when, for example, the "metric" in question sorts systems alphabetically by the system name as this produces perfectly consistent judgments regardless of the data used [22, 29]. However, we are interested in metrics that are strictly functions of a ranked list of items (i.e. system output) and a set of judged items (i.e. right answers). We are not interested in a "metric" that *knows* that (say) one ranked list is from Google and that the other is from Bing, and *uses this knowledge* to say which is better than the other.

The discriminative power method also provides a natural estimate of the performance difference (Δ) between two systems required to achieve statistical significance. This is done by recording, for every run pair, the Δ that corresponds to the borderline between significance and nonsignificance among the 1,000 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 bootstrap test, although other significance tests may be used just for computing discriminative power.

Other methods for evaluating evaluation metrics exist. The *swap method* proposed by Voorhees and Buckley [28] yields results similar to the discriminative power method but cannot directly examine the situation with the full topic set [19, 25]. The *maximum entropy method* [2, 17] can measure the informativeness of metrics but requires a mathematical derivation for each metric. Comparing the metrics with *user clicks* [5, 14] or with *user preferences* [23] should also be useful, and we expect these user-based approaches to complement our work.

[2]Clarke, Kolla and Vechtomova [9] have discussed the possibility of incorporating $Pr(i|q)$ into α-nDCG and NRBP.

[3]We call it D♯-measure because it "sharpens up" D-measure in terms of discriminative power, as we shall demonstrate later.

[4]Note that if we take the "top X% runs" from the run pool based on a metric, this may bias the metrics comparison results.

Table 1: Test collection statistics.

	(a) TREC 2009 Web Track Diversity Task Category A ("TR09DIV")	(b) NTCIR-6 CLIR Chinese ("NTCIR6C")
Documents	ClueWeb09 (approx. one billion Web pages) [6]	CIRB040r (901,446 Chinese news articles)
Topics	50 topics (12 ambiguous; 38 faceted) with 243 subtopics (177 informational; 66 navigational)	50
Intents	199 intents (i.e. subtopics with at least one relevant document). Max. #intents per topic: 6. Max. #intents per document: 5.	-
#relevant	across 50 topics: 4,942; across 199 intents: 6,499.	across 50 topics: 4,405 (1,807 $L1$, 1,519 $L2$ and 1,079 $L3$).

5. DATA

5.1 Adding Graded Relevance to TREC Data

Our experiments rely on the TREC2009 Web track diversity test collection with Category A runs [6], which we call TR09DIV. Some statistics are shown in Table 1(a). Unfortunately, TR09DIV contains neither the intent probabilities ($Pr(i|q)$) nor per-intent graded relevance assessments[5]. Hence, while this is a suitable situation for α-nDCG, it is difficult to demonstrate the advantages of other diversified IR metrics using TR09DIV.

Sakai *et al.* [20] compared D♯-measures, α-nDCG and nDCG-IA in terms of discriminative power and intuitiveness when the intent probability distribution is non-uniform. However, they did not use per-intent graded relevance. In contrast, this study fully utilises the capability of the diversified IR metrics to handle graded relevance, and considers both non-uniform and uniform distributions. To this end, we hired assessors to enrich the per-intent binary relevance data from TREC, as follows.

As Table 1(a) shows, we have 4,942 <topic, relevant document> pairs, or 6,499 <intent, relevant document> pairs. We assumed that all of these documents were judged as "partially relevant" by the TREC assessors[6]. Then, each intent-document pair was reassessed by two assessors: only this time, each assessor had a choice between "relevant" (the document fully satisfies the information need expressed in the subtopic field) and "partially relevant" (the document only partially satisfies the information need). The assessors used an assessment tool on which the TREC description and subtopic fields as well as the document content were displayed. To URLs that no longer exist, the default assessment "partially relevant" was given. Finally, we defined a relevance level for each document based on the three assessments (including the original one from TREC): $L3$ (two relevants and one partially relevant); $L2$ (one relevant and two partially relevants); and $L1$ (three partially relevants). We call the resultant data set TR09DIV+gr, where "gr" stands for (per-intent) graded relevance. The inter-assessor agreement between the new graded assessments is 69.9% (Cohen's kappa: .325). In this way, we obtained 1,173 $L1$, 1,959 $L2$ and 3,367 $L3$ documents across intents. Note that we did not re-examine any documents that were judged nonrelevant at TREC.

As for the intent probability distribution, we considered "Non-uniform" and "Uniform". Uniform means that all intents are equally likely, and this is the assumption currently used at the TREC Web diversity task. As for Non-uniform, we followed Sakai *et al.* [20]: for each topic with n intents, and assumed that the j-th intent has the probability $2^{n-j+1}/\sum_{k=1}^{n} 2^k$. Methods for estimating intent probabilities exist [1, 26], but our focus is on the inherent property of different diversified IR metrics *given* these probabilities.

5.2 Reducing the Diversity Test Collection

In addition to evaluating diversified IR metrics using TR09DIV+gr, we evaluated *traditional* graded-relevance IR metrics on which the diversified IR metrics are based. Sanderson *et al.* [23] treated each intent (i.e. subtopic) from TR09DIV as an independent topic to study the predictive power of traditional metrics. However, we avoided this approach because (a) This means that the traditional version of the test collection has many more topics than the original diversity version, and makes our traditional/diversity comparison rather difficult; and (b) The topic set thus constructed probably violates the i.i.d. assumption.

Instead, we constructed traditional test collections in two ways. The first method reduces TR09DIV+gr, by taking the *maximum* relevance level across intents for each topic-document pair. For example, if a document is $L1$-relevant to intent i_1 and $L3$-relevant to intent i_2 for a topic that has these two intents, then we treat this document as $L3$-relevant *to this topic* in the new collection. We call the new collection TR09DIV+gr2T, where "2T" means "(converted) to traditional." TR09DIV+gr2T has 601 $L1$, 1,328 $L2$ and 3,013 $L3$ documents across topics.

For comparison, we also constructed another traditional test collection using the original TR09DIV, not TR09DIV+gr. This was accomplished by simply defining graded relevance in terms of how many intents a document covers. The resultant collection, which we call TR09DIV2T, has 3,622 $L1$, 1,113 $L2$, 178 $L3$, 28 $L4$ and 1 $L5$ documents across topics. (As shown in Table 1(a), the maximum number of intents covered by a document in TR09DIV is 5.) Note that TR09DIV2T does not rely on our per-intent graded relevance data.

5.3 Another Data Set: NTCIR6C

In general, it is dangerous to try to draw strong conclusions from experiments that rely on a single test collection. We therefore conduct an additional set of traditional IR experiments using another data set. Our choice is the NTCIR-6 CLIR Chinese data (NTCIR6C) [12]: Table 1(b) shows its statistics. We selected NTCIR6C because (a) It is radically different from the TREC data discussed above in that it is from outside TREC and outside Web search (The NTCIR-6 task was a newspaper search task); and yet (b) It is similar to our TREC data in that it also has 50 topics and comes with relevance levels $L1$-$L3$.

6. EXPERIMENTS

6.1 Evaluating Traditional Search

Using the three traditional graded-relevance IR test collections (NTCIR6C, TR09DIV+gr2T and TR09DIV2T), we evaluated nDCG, Q, (n)GAP and (n)ERR in terms of discriminative power, for document cutoffs $l = 1000$ and $l = 10$. The cutoff of 1,000 represents classical TREC, while 10 represents the more recent shallow-depth evaluation practices as exemplified by the TREC Web tracks.

Figure 1 shows the *Achieved Significance Level (ASL) curves* [19] for nDCG, Q, (n)GAP and (n)ERR for each experimental condition. (Note that normalisation does not affect GAP and ERR when $l = 1000$.) The x axis represents the 190 run pairs sorted by ASL, and the y axis represents the ASL (i.e. p-value). Metrics whose graphs are closer to the origin are more discriminative than others, i.e. they are able to detect more significant differences. Table 2 cuts these graphs in the middle (horizontally) to compare the discrimi-

[5]For the TREC 2010 Web track, a kind of graded relevance was introduced to the ad hoc task but not to the diversity task.

[6]TREC binary relevance assessments are known to be "liberal," at least for early collections [27].

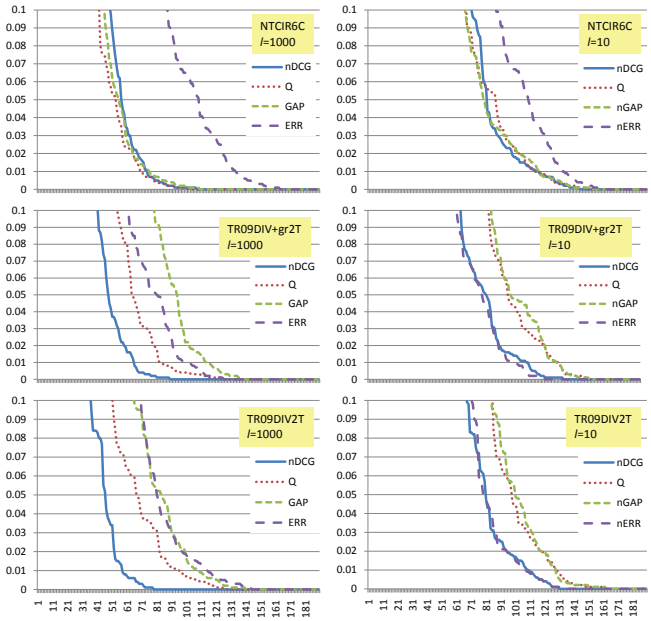

Figure 1: ASL curves for traditional metrics. The horizontal axes represent the 190 system pairs sorted by ASL; the vertical axes represent the ASL values.

native power at the .05 significance level. The table also shows the performance Δ that corresponds to a statistical significant difference. For example, with NTCIR6C at $l = 1000$, the discriminative power of Q-measure at the .05 level is $(137/190 =)72.1\%$, and if the Δ between two systems is 0.09 or larger, then that is usually statistically significant [19]. Figure 1 and Table 2 show that:

1. nDCG appears to be the most consistently discriminative metrics of all the traditional metrics for our data sets: Q does not perform as well as nDCG for the TREC data when $l = 10$; (n)GAP does not perform well for the TREC data when $l = 1000$ and when $l = 10$; (n)ERR doe not perform well for NTCIR6C, especially when $l = 1000$ as it suffers from its top-heaviness (i.e. virtually ignores all documents except for the very top ones).

2. Comparisons between $l = 1000$ and $l = 10$ show that, except for the top-heavy (n)ERR, using a small document cutoff (i.e. evaluating systems based on fewer data points) reduces discriminative power.

3. Comparisons across the three data sets with the same cutoff l show that the highest discriminative power achieved given 50 topics is similar: over 70% for $l = 1000$ and almost 60% for $l = 10$ at the .05 significance level. In particular, the results for TR09DIV+gr2T and TR09DIV2T are very similar, even though they have different number of relevance levels (up to $L3$ vs. up to $L5$). This suggests that the number of relevance levels and how they are obtained may not be major factors when comparing the discriminative power of metrics.

4. Normalisation improves the discriminative power of GAP, especially for NTCIR6C (Table 2(i), cutoff $l = 10$: the discriminative power goes up from 46.8% to 58.4%).

Table 2: Discriminative power of traditional metrics at $\alpha = .05$: Columns (a) and (b) show the discriminative power; (c) and (d) show the Δ required to achieve statistical significance.

cutoff $l = 1000$			cutoff $l = 10$		
	(a)	(b)		(c)	(d)
(i) NTCIR6C					
Q	72.1%	0.09	nGAP	58.4%	0.08
GAP	71.1%	0.07	nDCG	57.4%	0.11
nDCG	70.0%	0.11	Q	54.2%	0.09
ERR	42.6%	0.14	GAP	46.8%	0.04
			nERR	42.6%	0.16
			ERR	42.6%	0.15
(ii) TR09DIV+gr2T					
nDCG	74.7%	0.09	ERR	58.4%	0.17
Q	66.3%	0.05	nERR	57.9%	0.18
ERR	57.4%	0.15	nDCG	56.3%	0.11
GAP	50.5%	0.06	Q	48.4%	0.12
			nGAP	47.4%	0.13
			GAP	46.3%	0.02
(iii) TR09DIV2T					
nDCG	75.8%	0.07	nERR	58.4%	0.11
Q	64.7%	0.04	nDCG	57.9%	0.09
ERR	57.4%	0.04	ERR	50.0%	0.04
GAP	55.8%	0.05	Q	48.4%	0.08
			nGAP	47.4%	0.06
			GAP	43.2%	0.05

6.2 Evaluating Diversified Search

We now present our main results using TR09DIV+gr with per-intent graded relevance data. We evaluated I-rec, D(♯)-nDCG, D(♯)-Q, nDCG-IA, (n)GAP-IA, (n)ERR-IA as well as the official α-nDCG values at $l = 10$. As mentioned earlier, we considered both Non-uniform and Uniform intent probability distributions.

Figure 2 shows the ASL curves of the diversified IR metrics for four experimental conditions (two cutoffs $\times$ two intent probability distributions), and Table 3 compares the discriminative power at the .05 level. As α-nDCG and I-rec are not affected by intent probabilities, we show their results only in the Uniform results. Moreover, we do not consider I-rec (and therefore D♯-measures) when $l = 1000$ because I-rec is not useful with large document cutoffs: it would equal one most of the time. Also, we do not consider α-nDCG when $l = 1000$ as such values are not available from TREC. Figure 2 and Table 3 show that:

1. The Non-uniform and Uniform results are very similar. Thus, the intent probability distribution does not seem to have a major impact on discriminative power.

2. The most discriminative diversified IR metrics when $l = 1000$ are nDCG-IA and D-nDCG. The most discriminative when $l = 10$ are D♯-Q, α-nDCG, D♯-nDCG and I-rec.

3. The aforementioned "best" diversified IR metrics are at least as discriminative as the "best" traditional metrics: D-nDCG and nDCG-IA achieve well over 70% for $l = 1000$, while D♯-measures, α-nDCG and I-rec achieve well over 60% for $l = 10$ at the .05 level. (Hence, as in the traditional IR experiments, using a small document cutoff reduces discriminative power.)

4. In all four experimental conditions, (n)GAP-IA is substantially less discriminative than other metrics, and (n)ERR-IA is a middle-performer (it is as discriminative as nDCG-IA when $l = 10$, but less discriminative when $l = 1000$ due to its top-heaviness).

5. Normalisation improves the discriminative power of GAP-IA (Table 3, $l = 10$).

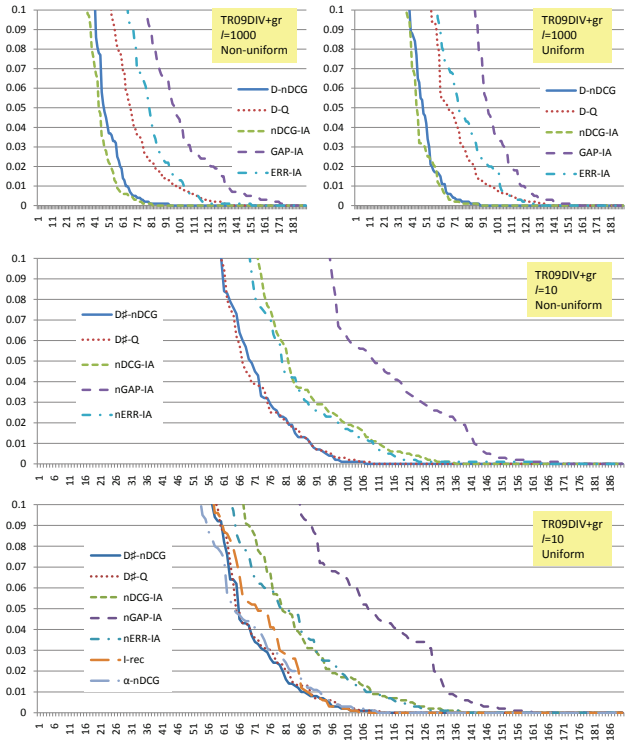

Figure 2: ASL curves for diversified IR metrics. The horizontal axes represent the 190 system pairs sorted by ASL; the vertical axes represent the ASL values.

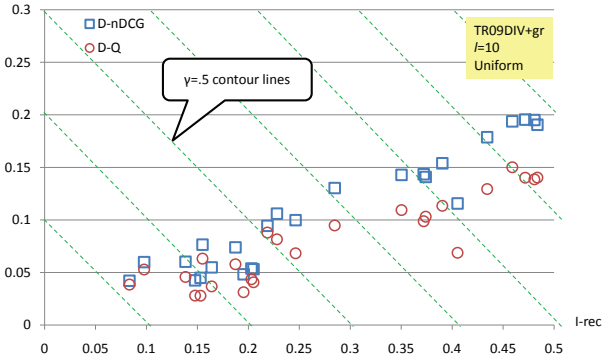

Figure 3: I-rec / D-measure graphs.

Note that the D♯-measures, which combine D-measures with I-rec, are slightly more discriminative than their components: for example, for the $l = 10$, Uniform case, the discriminative power of D-Q is 51.1%, that of I-rec is 62.6%, and that of D♯-Q is 66.3% at the .05 level (Table 3(ii)). Thus, D-measures and I-rec complement each other to achieve reliable evaluation results, by looking at a ranked list from two different (but not unrelated) angles. This generalises an observation by Sakai *et al.* [20].

Figure 3 plots all 25 runs for the TR09DIV+gr, $l = 10$, Uniform case with I-rec as the x axis and D-measure (i.e. D-nDCG or D-Q) as the y axis. The dotted lines represent the contour lines for a D♯-measure with $\gamma = .5$: if multiple runs lie on the same contour line, they are equally effective in terms of a D♯-measure. It can be observed that I-rec and D-measures are indeed already highly correlated with each other. A similar graph was shown by Sakai *et al.* [20] but they did not use per-intent graded relevance; the TREC 2009 overview paper [6] shows a similar graph for I-rec and the traditional *precision* metric.

Table 3: Discriminative power of diversified IR metrics at $\alpha = .05$: Columns (a) and (b) show the discriminative power; (c) and (d) show the Δ required to achieve statistical significance.

cutoff $l = 1000$			cutoff $l = 10$		
	(a)	(b)		(c)	(d)
(i) TR09DIV+gr; Non-uniform $P(i\|q)$					
nDCG-IA	77.4%	0.06	D♯-Q	65.3%	0.09
D-nDCG	75.3%	0.07	D♯-nDCG	63.7%	0.09
D-Q	61.1%	0.04	nERR-IA	58.4%	0.09
ERR-IA	58.4%	0.08	ERR-IA	57.9%	0.09
GAP-IA	48.9%	0.07	D-nDCG	57.4%	0.08
			nDCG-IA	57.4%	0.06
			D-Q	48.9%	0.09
			nGAP-IA	42.6%	0.07
			GAP-IA	24.2%	0.05
(ii) TR09DIV+gr; Uniform $P(i\|q)$					
nDCG-IA	76.8%	0.06	D♯-Q	66.3%	0.09
D-nDCG	75.3%	0.07	α-nDCG	66.3%	0.10
D-Q	63.2%	0.05	D♯-nDCG	65.8%	0.10
ERR-IA	61.6%	0.07	I-rec	62.6%	0.13
GAP-IA	50.5%	0.04	D-nDCG	60.5%	0.10
			nDCG-IA	58.4%	0.05
			nERR-IA	58.4%	0.08
			ERR-IA	57.9%	0.08
			D-Q	51.1%	0.10
			nGAP-IA	43.7%	0.04
			GAP-IA	35.3%	0.03

Table 4: τ and τ_{ap} (TR09DIV+gr; $l = 10$; Uniform).

	D♯-Q	D♯-nDCG	α-nDCG	nDCG-IA	nGAP-IA	nERR-IA
I-rec	.96/.97	.95/.86	.87/.74	.84/.84	.72/.70	.77/.63
D♯-Q	1/1	.97/.87	.87/.74	.88/.86	.74/.70	.79/.64
D♯-nDCG	-	1/1	.91/.87	.87/.76	.75/.62	.80/.76
α-nDCG	-	-	1/1	.84/.67	.74/.57	.90/.83
nDCG-IA	-	-	-	1/1	.85/.83	.86/.74
nGAP-IA	-	-	-	-	1/1	.78/.69

6.3 Examining Intuitiveness

Highly discriminative metrics, while desirable, may not necessarily measure what we want to measure. How do the different diversified IR metrics differ from one another, and which ones are more intuitive than others for the purpose of search result diversification?

Table 4 shows the Kendall's τ and τ_{ap} [30] values for different pairs of metrics, when the 20 runs are ranked in the TREC09DIV+gr, $l = 10$, Uniform setting. Kendall's 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 have an 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. Like τ, τ_{ap} lies between -1 and 1, but unlike τ, it is not symmetrical: one of the input rankings is taken as the gold standard. When the errors (i.e. pairwise item swaps with respect to the gold standard) are uniformly distributed over the ranking being examined, τ_{ap} is equivalent to τ. For example, the τ between I-rec and D♯-Q is .96, while the τ_{ap} between the same pair of metrics is .97 when D♯-Q is taken as the ground truth. The main message Table 4 conveys is that all of these metrics (including the simple I-rec) are reasonably correlated with one another.

Based on the bootstrap test results used for our discriminative power experiments, Table 5 shows the *agreement* between metrics, focussing on I-rec, D♯-measures, α-nDCG and nDCG-IA in the $l = 10$, Uniform setting. Let A and B denote the sets of sigfinicantly different run pairs at the .05 level according to two metrics, respectively. We define agreement as $|A \cap B|/|A \cup B|$. For example, the agreement between D♯-nDCG and α-nDCG is 86%, as $|A - B| = 9$, $|A \cap B| = 116$ and $|B - A| = 10$ for these two

Table 5: Agreement of significant differences at the .05 level (TR09DIV+gr; $l = 10$; Uniform).

	D♯-Q	D♯-nDCG	α-nDCG	nDCG-IA
I-rec	4/115/11 (86%)	4/115/10 (89%)	9/110/16 (81%)	23/96/15 (72%)
D♯-Q	-	1/125/0 (99%)	9/117/9 (87%)	21/105/6 (80%)
D♯-nDCG	-	-	9/116/10 (86%)	21/104/7 (79%)
α-nDCG	-	-	-	20/106/5 (81%)

metrics. That is, for nine run pairs, D♯-nDCG says that they are significantly different while α-nDCG says they are not; the situation is reversed for another set of ten run pairs[7]. One observation from this table is that the agreement between I-rec and nDCG-IA is relatively low (72%): this supports the arguments by Sakai *et al.* [20] and Clarke *et al.* [7] that IA metrics do not necessarily reward high intent recall, i.e. diversification. (See also the relatively low rank correlation values between I-rec and the IA metrics in Table 4.) On the other hand, it can be observed that the D♯-measures, α-nDCG and nDCG-IA agree with one another for around 80% of the time or more. (D♯-Q is virtually identical to D♯-nDCG: the agreement between them is 99%.) The important question is: when do they disagree? Hereafter, we focus our attention on D♯-nDCG, α-nDCG and nDCG-IA, as our experiments suggest that nDCG is the most reliable traditional IR metric when the document cutoff is small, and as these three metrics represent three different approaches to diversified IR evaluation.

To examine how D♯-nDCG, α-nDCG and nDCG-IA differ from the viewpoint of intuitiveness, we selected ten pairs of actual ranked lists from TREC 2009 Web track diversity runs as follows. First, from the aforementioned nine run pairs which were significantly different with D♯-nDCG but not with α-nDCG (Table 5), we obtained five pairs of ranked lists (i.e. run pairs for a particular topic) with the largest per-topic Δ's in terms of D♯-nDCG, under the constraint that there is a disagreement among D♯-nDCG, α-nDCG and nDCG-IA as to which run is better. We refer to these five cases as **A**-**E**, as shown in Table 6. For example, **Case A** in this table represents two runs watd3 and Sab9wtBfDiv for Topic 47 which has two intents. The middle column shows which of the top 10 documents retrieved by each run are relevant to which intent (where $\mathbf{i}$ and i indicate informational and navigational intents according to the TREC diversity topic file, respectively); the last three columns show the per-topic Δ's (e.g. performance of watd3 minus that of Sab9wtBfDiv), and the arrows indicate which run is rated higher with each metric. Similarly, from the ten run pairs which were significantly different with α-nDCG but not with D♯-nDCG shown in Table 5, five cases with the largest per-topic Δ's in terms of α-nDCG were selected. These are shown as cases **F**-**J** in Table 6. In short, these ten cases are the ones that contributed most to the discrepancy (in terms of statistical significance) between D♯-nDCG and α-nDCG. We shall closely examine these cases below from the viewpoint of intuitiveness. Note that these results are from the Uniform setting: Sakai *et al.* [20] have already compared the intuitiveness of D♯-nDCG, α-nDCG and nDCG-IA when the intent probability distribution is Non-uniform, a situation which α-nDCG does not handle.

We examined the ten cases shown in Table 6 and categorised the results into the following four classes. The arguments below are somewhat subjective, as the right balance between diversity and relevance is hard to define. Nevertheless, we believe that they are useful for understanding the diversity metrics.

[7] In theory, *conflicts* can also occur, where one metric says that run X significantly outperforms run Y while another metric says that run Y significantly outperforms X. There was no such case in our experiments.

6.3.1 *Only α-nDCG Prefers a Low-Relevance Run*

In **Case B**, we argue that α-nDCG is counterintuitive. Both twCSodpRBB and MSDiv1 completely failed to diversify: they cover the fifth intent $\mathbf{i}_5$ only. However, twCSodpRBB has only one relevant document (though at rank 1), while MSDiv1 has eight, all of which are $L3$-relevant. Since $\mathbf{i}_5$ is *informational*, MSDiv1 should probably be preferred, since the two runs are equally poor in terms of diversity (i.e. intent recall) but MSDiv1 has much better overall relevance. The rightmost columns of Table 6 show that D♯-nDCG and nDCG-IA agree with this intuition, while α-nDCG does not. This counterintuitiveness of α-nDCG is precisely because of α, which tends to ignore repetition of relevant documents for the same intent. It should be remembered that, unless the intent is purely navigational, providing the user with multiple documents that are relevant to the same intent does not necessarily imply redundancy in practice: different relevant documents may carry different pieces of information. Of course, a smaller value of α may remedy this particular situation, but how to appropriately set α in advance is an open question.

On the other hand, α-nDCG may be the most intuitive for **Case C**, which is similar to **Case B** in that both runs failed to diversify but different in that the intent involved is *navigational* ("Go to the Alexian Brothers Health System homepage"). Thus, even though D♯-nDCG and nDCG-IA prefer MSDiv1 which returned three documents that are $L3$-relevant to i_1, the second and the third $L3$-relevant documents may not be useful in practice. In particular, D♯-nDCG appears to favour MSDiv1 perhaps too much (the difference is over .15). Note that the raw nDCG is inherently suitable for evaluation with informative queries. We will discuss this point further in Section 6.3.5.

6.3.2 *α-nDCG and nDCG-IA Prefer a Non-diversified Run*

Next, let us discuss **Case A**, **Case F** and **Case H**, where only informational intents are involved and D♯-nDCG disagreed with α-nDCG and nDCG-IA. It can be observed that, in all three cases, α-nDCG and nDCG-IA prefer the non-diversified run, even though they are supposed to reward diversified ranked lists. In contrast, D♯-nDCG (with $\gamma = 0.5$) consistenly favours a more diversified run, due to its explict intent recall component. However, as we have discussed earlier, the right balance between relevance and diversity (which in the case of D♯-nDCG is represented by the γ parameter) is hard to define.

6.3.3 *Only α-nDCG Prefers a Poorly Diversifed Run*

Next, we discuss **Case D** and **Case E**, where only informational intents are involved and α-nDCG disagreed with D♯-nDCG and nDCG-IA. In both cases, α-nDCG prefers the less diversified run which missed the fourth intent (but nevertheless returned a document relevant to two intents at rank 1).

6.3.4 *Only nDCG-IA Prefers a Non-diversified Run*

In **Case G**, nDCG-IA is *clearly* counterintuitive. It can be observed that THUIR09FuClu has a document $L2$-relevant to $\mathbf{i_5}$ at rank 2, and one $L2$-relevant to $\mathbf{i_2}$ at rank 3. (These two intents are informational.) On the other hand, MSRABASE has only a document $L2$-relevant to i_3 at rank 2. (This intent is navigational.) Thus, since we are examining the uniform intent probability setting,

Table 6: Ten ranked list pairs from TR09DIV+gr, $l = 10$, Uniform. 2nd column: topic IDs (number of intents). 3rd column: run IDs. 4th column: number of intents covered by each run. 5th column: relevance levels for each intent at ranks 1-10. The rightmost columns: performance differences in terms of each metric; arrows point to the preferred run.

				document rank (**i**: informational intents; i: navigational intents)										Δ in D♯-nDCG	Δ in α-nDCG	Δ in nDCG-IA
				1	2	3	4	5	6	7	8	9	10			
A	47	watd3	1	$\mathbf{i_1}L3$	$\mathbf{i_1}L3$	$\mathbf{i_1}L3$			$\mathbf{i_1}L3$					**−0.209**	0.018	0.098
	(2)	Sab9wtBfDiv	2		$\mathbf{i_1}L3$					$\mathbf{i_2}L3$				⇓	⇑	⇑
					$\mathbf{i_2}L3$											
B	21	twCSodpRBB	1	$\mathbf{i_5}L3$										**−0.170**	0.023	−0.084
	(5)	MSDiv1	1			$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	$\mathbf{i_5}L3$	⇓	⇑	⇓
C	46	twCSodpRBB	1		i_1L3		i_1L3							**−0.156**	0.010	−0.031
	(3)	MSDiv1	1						i_1L3		i_1L3		i_1L3	⇓	⇑	⇓
D	26	watd3	2	$\mathbf{i_1}L2$		$\mathbf{i_1}L1$		$\mathbf{i_1}L2$					$\mathbf{i_1}L1$	**−0.141**	0.027	−0.025
	(4)			$\mathbf{i_3}L1$										⇓	⇑	⇓
		twCSodpRNB	3			$\mathbf{i_1}L1$				$\mathbf{i_1}L2$			$\mathbf{i_1}L2$			
						$\mathbf{i_3}L2$				$\mathbf{i_3}L3$			$\mathbf{i_3}L1$			
						$\mathbf{i_4}L1$				$\mathbf{i_4}L2$			$\mathbf{i_4}L2$			
E	26	watd1	2	$\mathbf{i_1}L2$	$\mathbf{i_1}L1$			$\mathbf{i_1}L1$		$\mathbf{i_1}L1$		$\mathbf{i_1}L2$		**−0.139**	0.042	−0.023
	(4)			$\mathbf{i_3}L1$										⇓	⇑	⇓
		twCSodpRNB	3			$\mathbf{i_1}L1$				$\mathbf{i_1}L2$			$\mathbf{i_1}L2$			
						$\mathbf{i_3}L2$				$\mathbf{i_3}L3$			$\mathbf{i_3}L1$			
						$\mathbf{i_4}L1$				$\mathbf{i_4}L2$			$\mathbf{i_4}L2$			
F	50	Sab9wtBfDiv	1	$\mathbf{i_3}L2$	$\mathbf{i_3}L2$			$\mathbf{i_3}L2$						−0.080	**0.168**	0.116
	(3)	spc	2					$\mathbf{i_1}L2$						⇓	⇑	⇑
								$\mathbf{i_2}L1$								
G	20	THUIR09FuClu	2		$\mathbf{i_5}L2$	$\mathbf{i_2}L2$								0.220	**0.150**	−0.092
	(4)	MSRABASE	1		i_3L2									⇑	⇑	⇓
H	14	Sab9wtBfDiv	1	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	$\mathbf{i_2}L1$	−0.102	**0.131**	0.028
	(4)	spc	2				$\mathbf{i_2}L1$						$\mathbf{i_2}L1$	⇓	⇑	⇑
													$\mathbf{i_4}L1$			
I	50	MSRABASE	3	$\mathbf{i_3}L2$	$\mathbf{i_3}L1$				$\mathbf{i_3}L2$	$\mathbf{i_1}L2$				0.236	**0.131**	−0.130
	(3)									$\mathbf{i_2}L1$				⇑	⇑	⇓
		THUIR09FuClu	1	$\mathbf{i_3}L2$	$\mathbf{i_3}L2$	$\mathbf{i_3}L2$	$\mathbf{i_3}L2$	$\mathbf{i_3}L2$		$\mathbf{i_3}L2$	$\mathbf{i_3}L2$	$\mathbf{i_3}L1$	$\mathbf{i_3}L1$			
J	38	Sab9wtBfDiv	2		$\mathbf{i_1}L2$		$\mathbf{i_3}L3$	$\mathbf{i_1}L2$			$\mathbf{i_1}L3$	$\mathbf{i_1}L2$		0.141	**0.130**	−0.019
	(3)	Sab9wtBDiv1	1			$\mathbf{i_1}L3$	$\mathbf{i_1}L2$	$\mathbf{i_1}L3$	$\mathbf{i_1}L3$	$\mathbf{i_1}L2$				⇑	⇑	⇓

THUIR09FuClu should definitely be preferred over MSRABASE, and both D♯-nDCG and α-nDCG satisfy this requirement.

The above counterintuitive behaviour of nDCG-IA arises from the inherent property of IA metrics, namely that high IA metric values can be achieved by doing extremely well for a single intent. In **Case G**, MSRABASE did extremely well for i_3: this intent only has one relevant document ($L2$), and the run returned this document at rank 2. The gain value for an $L2$-document is $2^2 - 1 = 3$, and therefore $nDCG_3 = (3/\log(2+1))/(3/\log(1+1)) = \log 2/\log 3 = .631$. Thus, nDCG-IA, averaged over four intents, is $.631/4 = .158$. Whereas, THUIR09FuClu achieves only $nDCG_2 = .119$ and $nDCG_5 = .144$ for the two intents and therefore its nDCG-IA is only $(.119 + .144)/4 = .066$.

Case I and **Case J** are similar to **Case G** in that only nDCG-IA prefers a run that failed to diversify. (All intents involved are informational.) We argue that nDCG-IA is rather counterintuitive for these cases as well, as MSRABASE covers three intents in **Case I** and Sab9wtBfDiv covers two in **Case J**.

6.3.5 *Intuitiveness Summary*

To sum up the above analysis: while the right balance between relevance and diversity is difficult to define, D♯-nDCG consistently prefers the more diversified run compared to α-nDCG and nDCG-IA. If we want diversity more than we want high relevance, then D♯-nDCG would be the clear winner, as it explitly incorporates intent recall. This is in contrast to α-nDCG which tries to encourage high intent recall by discouraging retrieval of "redundant" documents. As for IA metrics, we have demonstrated that they can be clearly counterintuitive, as high IA metric values can be achieved by retrieving highly relevant documents for one (major) intent.

The only case where D♯-nDCG may be less intuitive than α-nDCG is **Case C**, where the intent was navigational and therefore retrieving multiple relevant documents may not be practically useful. For navigational intents, it may be better to use graded-relevance versions of Reciprocal Rank such as ERR and P^+ [18][8].

Finally, recall that the above analysis used our Uniform results. Given a non-uniform intent probability distribution, α-nDCG can be more counterintuitive as it disregards the probabilities [20]. Also, recall that α-nDCG completely disregards local relevance levels (e.g. $L1$ vs $L3$ in Table 6).

7. CONCLUSIONS

To our knowledge, our work is the first to have studied the properties of different diversity evaluation metrics using per-intent graded relevance data. Moreover, our experiments are more extensive than similar studies that have been reported recently [7, 20]. Our main findings from the discriminative power experiments are:

- Our traditional IR and diversified IR experiments suggest that (n)GAP and (n)ERR and their intent-aware versions are *not* the most discriminative of metrics[9];
- D♯-measures, α-nDCG and intent recall appear to be the most discriminative metrics for shallow-depth diversified IR eval-

[8]Q-measure, P^+ and ERR are all members of a family of metrics called *Normalised Cumulative Utiliy* (NCU) [21]. An NCU is defined in terms of a user's stopping probability distribution over ranks and a utility function at a given rank. Q-measure uses a uniform distribution over all relevant documents; P^+ uses a uniform distribution over all relevant documents retrieved within top r_p, where r_p is the rank of one of the most relevant documents in the ranked list. Whereas, ERR's stopping probability at a given rank depends on the relevance of previously seen documents.

[9]Recall, however, that we have only examined a version of GAP which uses a flat probability distribution over the relevance levels, a setting recommended by Robertson, Kanoulas and Yilmaz [17].

Table 7: Summary: properties of diversified IR metrics.

	α-nDCG	IA metrics	D♯-measures
Intent probabilities	NO	YES	YES
Pay attention to minor intents	YES	NO	YES
Per-intent graded relevance	NO	YES	YES
Max value guaranteed to be 1	NO	NO	YES
High discriminative power	YES	NO	YES

uation. Moreover, they are at least as discriminative as the most discriminative traditional metrics (e.g. nDCG).

The second finding suggests that diversified IR experiments with a given number of topics can be as reliable as traditional IR experiments with the same number of topics, provided that the aforementioned discriminative metrics are used. This is good news for IR test collection builders and users.

Moreover, our analysis showed that, while different diversified IR metrics are generally highly correlated with one another, D♯-nDCG is more intuitive than α-nDCG and nDCG-IA at least when high diversity is considered more important than high relevance.

Table 7 summarises the properties of diversified IR metrics examined in this study. The original α-nDCG (as used in TREC) can handle neither intent probabilities nor per-intent graded relevance; IA metrics tend to ignore minor intents; α-nDCG and IA metrics have normalisation issues. In contrast, D-measures range fully between 0 and 1, and so do D♯-measures provided that the measurement depth l is not smaller than the number of intents. In addition to these inherent differences, our present study showed that IA metrics have relatively low discriminative power, and that D♯-measures have strengths in terms of intuitiveness. (The latter observation is not included in Table 7 as it is somewhat subjective.) It is probably fair to say that the D♯-measures are promising for diversified IR evaluation. A practical recommendation for diversified IR evaluation would be to plot I-rec against D-nDCG as we have shown in Figure 3 and to discuss the contour lines that represent D♯-nDCG.

Our results on diversified IR metrics, however, rely solely on TR09DIV+gr (just as other studies [7, 20, 23] relied solely on TR09DIV). As future work, we plan to construct more diversity test collections (at the NTCIR "INTENT" task[10]) and strengthen our conclusions. We also plan to explore related questions such as: (1) how to seamlessly evaluate diversity and relevance for navigational and informational queries; and (2) how to evaluate diversity across verticals in aggregated search and across queries in a session, and formulate a unified, general framework for diversity evaluation.

8. REFERENCES

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

[2] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proceedings of ACM SIGIR 2005*, pages 27–34, 2005.

[3] C. Brandt, T. Joachims, Y. Yue, and J. Bank. Dynamic ranked retrieval. In *Proceedings of WSDM 2011*, pages 247–256, 2011.

[4] 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.

[5] 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.

[6] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *Proceedings of TREC 2009*, 2010.

[7] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of ACM WSDM 2011*, 2011.

[10] http://www.thuir.org/intent/ntcir9/

[8] 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.

[9] 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.

[10] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of ACM WSDM 2011*, 2011.

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

[12] K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, and H.-H. Chen. Overview of CLIR task at the sixth NTCIR workshop. In *Proceedings of NTCIR-6*, 2007.

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

[14] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *Proceedings of ACM SIGIR 2010*, pages 667–674, 2010.

[15] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of ACM WWW 2010*, pages 781–790, 2010.

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

[17] S. E. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *Proceedings of ACM SIGIR 2010*, pages 603–610, 2010.

[18] T. Sakai. Bootstrap-based comparisons of IR metrics for finding one relevant document. In *Proceedings of AIRS 2006 (LNCS 4182)*, pages 374–389, 2006.

[19] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532, 2006.

[20] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin. Simple evaluation metrics for diversified search results. In *Proceedings of EVIA 2010*, pages 42–50, 2010.

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

[22] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4:247–375, 2010.

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

[24] R. L. Santos, C. Macdonald, and I. Ounis. Selectively diversifying search results. In *Proceedings of ACM CIKM 2010*, 2010.

[25] I. Soboroff. Test collection diagnosis and treatment. In *Proceedings of EVIA 2010*, pages 34–41, 2010.

[26] 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*, pages 51–59, 2010.

[27] E. Sormunen. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of ACM SIGIR 2002*, pages 324–330, 2002.

[28] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of ACM SIGIR 2002*, pages 316–323, 2002.

[29] W. Webber, A. Moffat, and J. Zobel. The effect of pooling and evaluation depth on metric stability. In *Proceedings of EVIA 2010*, pages 7–15, 2010.

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

[31] 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.