

Evaluation with Informational and Navigational Intents

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

ABSTRACT

Given an ambiguous or underspecified query, search result diversification aims at accomodating different user intents within a single “entry-point” result page. However, some intents are informational, for which many relevant pages may help, while others are navigational, for which only one web page is required. We propose new evaluation metrics for search result diversification that considers this distinction, as well as a simple method for comparing the intuitiveness of a given pair of metrics quantitatively. Our main experimental findings are: (a) In terms of discriminative power which reflects statistical reliability, the proposed metrics, $D_{\#}$ -nDCG and $P+Q_{\#}$, are comparable to intent recall and $D_{\#}$ -nDCG, and possibly superior to α -nDCG; (b) In terms of preference agreement with intent recall, $P+Q_{\#}$ is superior to other diversity metrics and therefore may be the most intuitive as a metric that emphasises diversity; and (c) In terms of preference agreement with effective precision, $D_{\#}$ -nDCG is superior to other diversity metrics and therefore may be the most intuitive as a metric that emphasises relevance. Moreover, $D_{\#}$ -nDCG may be the most intuitive as a metric that considers both diversity and relevance. In addition, we demonstrate that the randomised Tukey’s Honestly Significant Differences test that takes the entire set of available runs into account is substantially more conservative than the paired bootstrap test that only considers one run pair at a time, and therefore recommend the former approach for significance testing when a set of runs is available for evaluation.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

diversification, evaluation, intents, metrics, novelty, redundancy

1. INTRODUCTION

Search queries are often *ambiguous* (e.g., “office” can be a workplace or a software) and/or *underspecified* (e.g., “harry potter” can be a book, a film or the main character) [9]. To accommodate such different user needs (or user *intents*) given a query, research in *search result diversification* has received much attention recently (e.g., [1, 10, 14, 15, 24]). TREC began a diversity task in the

Web track in 2009 [6]¹, and NTCIR concluded its first INTENT task for mining intents and selectively diversifying search results in 2011 [26]². These tasks evaluate a flat list of possibly diversified but relevant documents, although other approaches such as dynamic presentation [2] may also be useful.

The main challenge in diversity evaluation is the balancing between diversity and relevance. That is, we want search engines to cover as many intents as possible in the first Search Engine Result Page (SERP), but we also want as many relevant documents as possible. Moreover, if we know that some intents for a given query are more *likely* than others, we might want to allocate more space within the SERP to the popular intents. Furthermore, we probably want documents that are *highly* relevant to each intent rather than those that are *partially* relevant. We need “good” evaluation metrics that reflect these requirements, in order to achieve the goal of providing a single “entry-point” SERP that is useful to as many users as possible.

In light of the above considerations, Sakai and Song [21] conducted an extensive study of different diversity metrics in terms of *discriminative power* [18] and intuitiveness, given the premises that *intent probabilities* and *per-intent graded relevance assessments* are available with the diversity test collection. Discriminative power is the proportion of statistically significant differences one can get out of a given experimental environment and therefore a measure of how reliable a metric is. (Details will be given in Section 4.1.) Sakai and Song also discussed intuitiveness by manually examining pairs of ranked lists, and showed that a family of metrics called *$D_{\#}$ -measures* [21] have several advantages over α -nDCG [8] and *Intent-Aware* (IA) metrics [1]. More specifically, they highlighted the following limitations of α -nDCG and IA metrics:

1. α -nDCG can handle neither intent probabilities nor per-intent graded relevance (although intent probabilities were later incorporated [7, 9]).
2. IA metrics can be clearly counterintuitive at times. They also tend to reward non-diversified systems that focus on popular intents [7], and have relatively low discriminative power.
3. α -nDCG and IA metrics are not guaranteed to lie fully between 0 and 1.

Sakai and Song [21] also showed that *Expected Reciprocal Rank* (ERR) [5] and *Graded Average Precision* (GAP) [16] have low discriminative power both as traditional IR metrics and as IA metrics for diversity evaluation.

¹<http://plg.uwaterloo.ca/~trecweb/>

²<http://research.microsoft.com/en-us/people/tesakai/intent2.aspx>

While the three problems mentioned above do not apply to $D_{\#}$ -measures, the manual analysis by Sakai and Song [21] suggested that $D_{\#}$ - $nDCG$, a member of the $D_{\#}$ -measure family, may be less intuitive than α - $nDCG$ when the intents are *navigational*. Conversely, α - $nDCG$ seemed less intuitive than $D_{\#}$ - $nDCG$ when the intents are *informational*. The original definitions of navigational and informational intents by Broder [3] are:

Navigational The immediate intent is to reach a particular site.

Informational The intent is to acquire some information assumed to be present on one or more web pages.

Thus, according to these definitions, there is basically only one web page that the user wants to see when the intent is navigational, while the user may be happy to see many relevant pages (minus duplicate information) when the intent is informational. α - $nDCG$ works well for navigational intents precisely because of its α , which discourages retrieval of multiple relevant documents for each intent. Whereas, $D_{\#}$ - $nDCG$ works well for informational intents, precisely because $nDCG$ (*normalised Discounted Cumulative Gain*) [12] was designed to *cumulate* pieces of information across multiple relevant documents. According to a study by Jansen, Booth and Spink [11], over 80% of their Dogpile metasearch queries were informational, and about 10% were navigational, although multi-intent queries were outside the scope of their study.

The objective of this paper is to explore ways to incorporate the explicit knowledge of informational and navigational intents into diversity evaluation, and to design diversity metrics that are more intuitive than $D_{\#}$ -measures and α - $nDCG$. We propose new diversity evaluation metrics called $DIN_{\#}$ -measures and $P+Q_{\#}$, as well as a simple method for comparing the intuitiveness of a given pair of metrics quantitatively. Our main experimental findings are:

- In terms of *discriminative power* [18] which reflects statistical reliability, the proposed metrics, $DIN_{\#}$ - $nDCG$ and $P+Q_{\#}$, are comparable to intent recall and $D_{\#}$ - $nDCG$, and possibly superior to α - $nDCG$;
- In terms of *preference agreement* (described in Section 4.2) with intent recall, $P+Q_{\#}$ is superior to other diversity metrics and therefore may be the most intuitive as a metric that emphasises diversity; and
- In terms of preference agreement with *effective precision* (also described in Section 4.2), $DIN_{\#}$ - $nDCG$ is superior to other diversity metrics and therefore may be the most intuitive as a metric that emphasises relevance. Moreover, $DIN_{\#}$ - $nDCG$ may be the most intuitive as a metric that considers both diversity and relevance.

In addition, we demonstrate that the *randomised Tukey's Honestly Significant Differences test* [4] that takes the entire set of available runs into account is substantially more conservative than the *paired bootstrap test* [18] that only considers one run pair at a time, and therefore recommend the former approach for significance testing when a set of runs is available for evaluation.

The remainder of this paper is organised as follows. Section 2 discusses previous work related to this study and defines existing diversity metrics. Section 3 defines our proposed metrics, and Section 4 describes how we evaluate diversity evaluation metrics in terms of discriminative power and intuitiveness. Section 5 describes our experiments and reports on discriminative power and intuitiveness results. Finally, Section 6 concludes this paper.

2. PREVIOUS WORK

This section summarises prior art related to this study. Section 2.1 first defines some traditional graded-relevance IR metrics on top of which diversified IR metrics have been designed. Section 2.2 defines these existing diversity metrics, namely, α - $nDCG$, IA metrics and $D_{\#}$ -measures. Section 2.3 summarises previous findings from comparing different diversity metrics.

2.1 Traditional Metrics

We first define a popular version of $nDCG$. Let $g(r)$ denote the *gain value* at rank r in a system's ranked list. Following a popular practice, we let $g(r) = 7$ if the document at r is *highly relevant* ($L3$), $g(r) = 3$ if it is *relevant* ($L2$), and $g(r) = 1$ if it is *partially relevant* ($L1$). Otherwise $g(r) = 0$. The *cumulative gain* at rank r is defined as $cg(r) = \sum_{k=1}^r g(k)$. Also, let $g^*(r)$ and $cg^*(r)$ denote the (cumulative) gain at rank r in an *ideal ranked list*, obtained by listing up all relevant documents in descending order of relevance levels. $nDCG$ at document cutoff l can be defined as:

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

Let $J(r) = 0$ if a document at rank r is nonrelevant to the query and $J(r) = 1$ otherwise. Let $C(r) = \sum_{k=1}^r J(k)$. Then the *blended ratio* at rank r , a graded-relevance version of precision, is defined as:

$$BR(r) = \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} \quad (2)$$

where $\beta (\geq 0)$ is a user persistence parameter which is set to 1 throughout this study. Then Q -measure [18, 20] is defined as:

$$Q\text{-measure} = \frac{1}{R} \sum_{r=1}^L J(r) BR(R) \quad (3)$$

where L is the size of the ranked list. Note that $\beta = 0$ reduces Q -measure to Average Precision. Since we are interested in evaluation with a small document cutoff to evaluate the *first* SERP, we use a document-cutoff version of Q -measure, $Q@l$, which replaces the R with $\min(l, R)$ and the L with l in Eq. 3 to ensure that the maximum value achievable is 1.

As can be seen, both $nDCG$ and Q are defined based on *cumulating* gains discounted by ranks, and are inherently suitable for *informational* queries where more relevant documents means better user satisfaction. But there also exist metrics that are more suitable for *navigational* queries, for which obtaining exactly one (highly) relevant document is sufficient. ERR and P^+ [17] are examples of such metrics. ERR assumes that the user is dissatisfied with documents from ranks 1 to $r-1$ and is finally satisfied with one at rank r , and that the satisfaction probabilities are proportional to the gain values. Whereas, P^+ assumes that the user stops examining the ranked list at the *preferred rank* (rp), which contains one of the most relevant documents within the ranked list and is closest to the top of the list. In this paper, as we are interested in evaluation with a small document cutoff, we define rp after truncating the ranked list at the cutoff³.

Formally, P^+ is defined as [17]:

$$P^+ = \frac{1}{C(rp)} \sum_{r=1}^{rp} J(r) BR(R) \quad (4)$$

³Suppose that the cutoff $l = 10$, and the system output has an $L1$ -relevant document at rank 1, and two $L2$ -relevant documents at ranks 5 and 10, and one $L3$ -relevant document at rank 20. Then, in our setting, $rp = 5$ as we ignore the document at rank 20.

if there is at least one relevant document in the (truncated) ranked list, and $P^+ = 0$ otherwise.

Sakai [17] showed that metrics for navigational topics (such as P^+) generally have lower discriminative power than those for informational topics (such as Q) as the former generally rely on fewer data points, i.e., retrieved documents *treated* as relevant. Similarly, as was mentioned earlier, Sakai and Song [21] reported somewhat negative results for ERR in terms of discriminative power.

Q-measure, P^+ and ERR can be seen as members of the *Normalised Cumulative Utility* (NCU) metrics family [20]. An NCU metric is defined as a combination of the user’s stopping probability distribution across document ranks and a utility function given a particular stopping rank. Q-measure’s probability distribution is uniform across *all* relevant documents; that of P^+ is uniform across all relevant documents retrieved between ranks 1 and rp . Both metrics measure the utility by means of the aforementioned blended ratio. Whereas, both ERR and a *rank-biased* version of NCU [20] use stopping probabilities that depend on the number of relevant documents previously seen.

2.2 Diversity Metrics

α -nDCG is an extension of nDCG towards diversity evaluation. It views both query intents and documents as sets of *nuggets*. The main idea is to discount the gains according to “nuggets already seen” before discounting by ranks. The strength of the novelty-biased discount is controlled by α (which is set of 0.5 throughout this paper as we use the official α -nDCG values from the TREC 2009 Web track). Formally, let $J_n(r) = 1$ if a document at rank r is relevant to the n -th nugget and 0 otherwise; let $C_n(r) = \sum_{k=1}^r J_n(k)$, i.e., the number of documents observed within top r that contained the n -th nugget. Then the *novelty-biased gain* is defined as $NG(r) = \sum_{n=1}^m J_n(r)(1 - \alpha)^{C_n(r-1)}$, where m is the total number of nuggets for the query. α -nDCG is defined by replacing the raw gain values in Eq. 1 with the novelty-biased gains.

Unlike the IA metrics and the $D(\#)$ -measures discussed below, the original α -nDCG [8] can handle neither intent likelihood nor per-intent graded relevance. Leenanupab, Zuccon and Jose [13] have proposed to adjust the value of α per topic, which may improve the intuitiveness of α -nDCG. However, this approach does not change the above two limitations.

Given the intent probabilities $P(i|q)$ for intent i and query q , where $\sum_i Pr(i|q) = 1$, as well as per-intent graded relevance assessments, an IA version of a given metric M is given by

$$M-IA = \sum_i Pr(i|q)M_i \quad (5)$$

where M_i is the per-intent (or *local*) version of metric M . For example, *nDCG-IA* is computed as follows: (1) Define an ideal ranked list *for each intent*; (2) For each intent, compare the system output with the local ideal list and compute the local nDCG ($nDCG_i$); (3) Finally, apply Eq. 5.

Both α -nDCG and IA metrics are *imperfect* metrics in that they are not guaranteed to range between 0 and 1: computing an ideal list for α -nDCG is NP-complete; it is generally not possible for a single system output to be ideal for all intents at the same time.

We now define the D-measures, which are free from the aforementioned limitations of α -nDCG and the IA metrics. Given the intent probabilities $Pr(i|q)$ and per-intent graded relevance assessments, where $g_i(r)$ is the gain value for document at rank r for intent i , we first define the *global gain* at rank r as:

$$GG(r) = \sum_i Pr(i|q)g_i(r). \quad (6)$$

We then define a single ideal list (in contrast to the IA metrics which define an ideal list for every intent) by sorting all relevant documents by the global gain, and denote the ideal global gain at rank r by $GG^*(r)$. Finally, by replacing the raw gains of metrics such as nDCG and Q-measure with the global gains, D-measures (D-nDCG, D-Q etc.) can be computed. Note that there is no NP-complete problem involved here.

Sakai and Song [21] proposed to plot D-measures against *intent recall* (a.k.a. *subtopic recall* [28], the proportion of intents covered by a ranked list) to visualise the trade-off between relevance and diversity. In addition, to obtain a single-value metric, they proposed to compute the $D\#$ -measures in addition:

$$D\#-measure = \gamma I-rec + (1 - \gamma) D-measure \quad (7)$$

where γ is a parameter. Throughout this paper, we let $\gamma = 0.5$: intent recall and D-nDCG/Q are highly correlated with each other and therefore $D\#$ -nDCG/Q are not so sensitive to the choice of γ [21].

2.3 Comparing Diversity Metrics

To date, there are only a few studies that compared the reliability and usefulness of different diversity metrics.

Clarke *et al.* [7] compared diversity metrics including α -nDCG, a similar metric called *Novelty- and Rank-Biased Precision* (NRBP) and an IA version of ERR (ERR-IA) in terms of discriminative power. Somewhat surprisingly, their results suggested that intent recall, a simple set-based diversity metric, is more discriminative than others. However, their experiments were limited to *uniform* intent probabilities and *binary* per-intent graded relevance assessments from the TREC 2009 Web track test collection [6].

Sakai and Song [21] compared $D(\#)$ -measures with α -nDCG and a variety of IA metrics including ERR-IA, using uniform and nonuniform intent probabilities and *graded* per-intent relevance assessments added to the TREC 2009 Web collection. They compared the metrics in terms of discriminative power *and* intuitiveness: their results suggested that $D\#$ -measures are the most promising diversity metrics among the existing ones. Also, as was mentioned earlier, their intuitiveness analysis suggested that while α -nDCG may sometimes be more intuitive than other metrics for navigational intents, $D\#$ -measures may be more intuitive for informational intents, which is the main motivation of this study. Moreover, as Sakai and Song’s intuitiveness analysis was somewhat subjective and anecdotal, we propose a simple method for quantifying the relative intuitiveness of diversity metrics in this present study.

Using the Amazon Mechanical Turk framework and the TREC 2009 Web collection with *binary* relevance assessments, Sanderson *et al.* [23] examined the *predictive power* of diversity metrics such as α -nDCG: if a metric prefers one ranked list over another, does the user also prefer the same list? While our method for quantifying the relative intuitiveness of diversity metrics was partially inspired by the side-by-side approach of Sanderson *et al.*, their work and ours fundamentally differ in the following aspects: (1) While Sanderson *et al.* treated each subtopic (i.e., intent) as an independent topic to examine the relationship between user preferences and metric preferences, we aim to measure the intuitiveness of metrics with respect to the entire (ambiguous or underspecified) topic in terms of diversity and relevance; (2) While Sanderson *et al.* used the Mechanical Turkers, we use very simple evaluation metrics that represent diversity or relevance as the gold standard in order to quantify the intuitiveness. Sanderson *et al.* found that intent recall (called “cluster recall” in their paper) is as effective as other diversity metrics in predicting user preferences, despite its simplicity. They also reported that diversity metrics agreed well with user

preferences especially for navigational (sub)topics, although their analysis relied on only 18 navigational subtopics.

3. PROPOSED METRICS

This section proposes new diversity metrics that rely on the explicit knowledge on whether an intent is informational or navigational.

3.1 DIN-measures and $DIN_{\#}$ -measures

Our first proposal, DIN-measures⁴, are identical to D-measures in the way the globally ideal ranked list is defined. The only difference is that systems do not receive any credit for returning multiple relevant documents for each *navigational* intent. For example, consider a ranked list shown in Figure 1 for a query with exactly one informational intent i and exactly one navigational intent j . Suppose that, as the figure shows, the document at rank 1 is $L1$ -relevant to i , the document at rank 2 is $L3$ -relevant to i and $L1$ -relevant to j , and so on. While existing diversity measures such as α -nDCG and D-nDCG consider the document at rank 4 as relevant to j , DIN-measures treats this document as nonrelevant to j because a relevant document has already been found at rank 2 for this navigational intent. (As this example shows, even navigational intents may have multiple relevant documents in the test collection.) Note that this is similar to how the binary-relevance *Reciprocal Rank* evaluates a ranked list: only the first relevant document matters.

Formally, let $\{i\}$ and $\{j\}$ denote the sets of informational and navigational intents for query q , and let $isnew_j(r) = 1$ if there is no document relevant to the navigational intent j between ranks 1 and $r - 1$, and $isnew_j(r) = 0$ otherwise. We redefine the Global Gain as:

$$GG^{DIN}(r) = \sum_i Pr(i|q)g_i(r) + \sum_j isnew_j(r)Pr(j|q)g_j(r). \quad (8)$$

This should be compared with the original Global Gain (Eq. 6) which does not distinguish between informational and navigational intents. It can be observed that GG^{DIN} simply ignores redundant relevant documents for navigational intents.

Now, *DIN-nDCG*, for example, can be defined as:

$$DIN-nDCG@l = \frac{\sum_{r=1}^l GG^{DIN}(r)/\log(r+1)}{\sum_{r=1}^l GG^*(r)/\log(r+1)}. \quad (9)$$

Similarly, *DIN-Q* can be defined as:

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

where

$$DIN-BR(r) = \frac{C(r) + \beta \sum_{k=1}^r GG^{DIN}(k)}{r + \beta \sum_{k=1}^r GG^*(k)}. \quad (11)$$

Note that only the system's global gains (numerators in Eqs. 9 and 11) have been modified, and the ideal global gains (denominators) remain unchanged. This means that, unlike D-measures, the maximum possible value of a DIN-measure may be less than one. We regard this as a cost of improving the intuitiveness of diversity metrics while keeping them simple.

Just like D-measures, DIN-measures can be combined with intent recall (or *I-rec*) to boost diversity relative to relevance (Recall Eq. 7). We call the resultant metrics $DIN_{\#}$ -metrics. In this paper,

⁴DIN stands for: Diversification for Informational and Navigational intents.

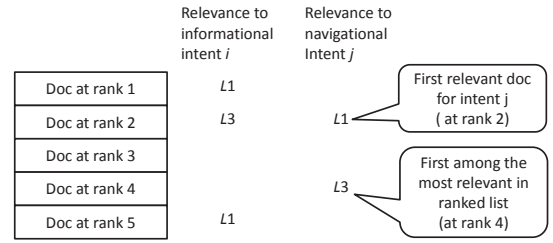


Figure 1: An example ranked list for a query with one informational intent and one navigational intent.

we examine $DIN_{\#}$ -nDCG and $DIN_{\#}$ -Q: the latter uses the cutoff version of Q-measure as was described in Section 2.1.

3.2 P+Q and $P+Q_{\#}$

Our second proposal is to extend the IA approach of Agrawal *et al.* [1], so that two different metrics are used for informational and navigational intents, respectively. A natural choice would be to use two metrics that share a similar user model: in this paper, we use $Q@l$ for informational intents, and P^+ for navigational intents, and call the resultant metric $P+Q$:

$$P+Q@l = \sum_i Pr(i|q)Q_i@l + \sum_j Pr(j|q)P_j^+. \quad (12)$$

Here, for example, $Q_i@l$ means $Q@l$ computed for intent i based on an ideal list defined particularly for this intent. Recall also that, in this paper, the preferred rank rp_j for each P_j^+ is defined after truncating the ranked list at l , and therefore $rp_j \leq l$ holds (See Section 2.1).

Let us go back to Figure 1: $P+Q$ is computed for this example as follows. For the informational intent i , $Q@l$ is computed by taking the relevant documents at ranks 1, 2 and 5 into account: recall that Q assumes that the user is equally likely to stop examining the ranked list at any of these three ranks. Whereas, for the navigational intent j , we first determine rp : in this example, $rp = 4$ (not 2), because the highest relevance level found in the ranked list is $L3$ and the document at rank 4 is the first document whose relevance level is $L3$. Then, P^+ for j is computed: recall that it assumes that the user is equally likely to stop examining the ranked list at ranks 2 and 4. Finally, the value of Q and P^+ are combined by taking the intent probabilities into account. Note that, in this particular example, P^+ is the same as Q for intent j and therefore $P+Q$ is the same as Q -IA, the Intent-Aware version of Q . Whereas, if the document at rank 2 in Figure 1 was (say) $L3$ -relevant for j , then the document at rank 4 would be ignored and $P+Q$ would be less than Q -IA.

Just like the IA metrics, the maximum value of $P+Q$ is usually below 1: a single system output is almost never ideal for all intents at the same time. Again, we regard this as a cost of improving the intuitiveness of diversity metrics while keeping them simple.

Furthermore, we consider combining $P+Q$ with *I-rec* to emphasise diversity, and call the resultant metric $P+Q_{\#}$. Note that Sakai and Song [21] did not consider the combination of IA metrics with *I-rec*, although it is possible⁵.

⁵Combining α -nDCG and *I-rec* in a similar way would be redundant, as α -nDCG already has a mechanism for emphasising diversity, namely the parameter α .

```

 $\mathbf{z} = (z_1, \dots, z_N)$  where  $z_i = M(t_{T,i}, r_1) - M(t_{T,i}, r_2)$ ;
 $t(\mathbf{z}) = \frac{\bar{z}}{\bar{\sigma}/\sqrt{N}}$  where  $\bar{z}$  and  $\bar{\sigma}$  are mean and standard deviation of  $\mathbf{z}$ ;
 $\mathbf{w} = (z_1 - \bar{z}, \dots, z_N - \bar{z})$ ;
 $count = 0$ ;
for  $b = 1$  to  $B$  do {
   $\mathbf{w}^{*b}$  = bootstrap sample of size  $N$ 
    obtained by sampling with replacement from  $\mathbf{w}$ ;
   $t(\mathbf{w}^{*b}) = \frac{\bar{w}^{*b}}{\bar{\sigma}^{*b}/\sqrt{N}}$  where  $\bar{w}^{*b}$  and  $\bar{\sigma}^{*b}$  are
    mean and standard deviation of  $\mathbf{w}^{*b}$ ;
  if  $(|t(\mathbf{w}^{*b})| \geq |t(\mathbf{z})|)$   $count++$ ;
}
 $ASL = count/B$ ;

```

Figure 2: Algorithm for obtaining the Achieved Significance Level with the two-sided, paired bootstrap test given two runs r_1 and r_2 , Topic Set T ($|T| = N$) and Metric M [18].

```

foreach pair of runs  $(r_1, r_2)$  do
  if  $(|t(\mathbf{w})^{*b'}|$  is the  $B\alpha$ -th largest value in  $\{|t(\mathbf{w})^{*b}\}$ )
     $\Delta_\alpha(r_1, r_2) = |\bar{w}^{*b'}|$ ;
 $\Delta_\alpha = \max_{i,j} \Delta_\alpha(r_i, r_j)$ ;

```

Figure 3: Algorithm for estimating the performance Δ required for obtaining a significant difference at α with the paired bootstrap test [18].

4. EVALUATING EVALUATION METRICS

This section describes two methods for comparing the “goodness” of diversity metrics: *discriminative power* [18], which represents the statistical reliability of a metric, and *intuitiveness test*, which is our new proposal.

4.1 Discriminative Power

Given a test collection with a set of runs, discriminative power is measured by conducting a statistical significance test for every pair of runs and counting the number of significant differences. In this paper, we use two different significance tests that rely on computer power and thereby require fewer assumptions than classical tests such as the t -test. The first is the *paired bootstrap test* which was the significance test originally used for measuring discriminative power [18]. The second is the *randomised version of Tukey’s Honestly Significant Differences (HSD) test* [4].

The bootstrap test is conducted for every run pair independently. That is, the statistical significance at α (i.e., Type I error probability: note that this is unrelated to α -nDCG) for a run pair is tested without taking the other runs into consideration. However, pairwise tests conducted in this fashion for k run pairs inevitably results in the *family-wise error rate* of $1 - (1 - \alpha)^k$ [4]: this is the probability of detecting at least one significant difference for a pair of runs that are in fact no different from each other. Note that this problem applies to *all* pairwise significance tests.

In contrast, the randomised Tukey’s HSD test takes the entire set of runs into account to judge whether each run pair is significantly different or not. Thus this test is naturally more *conservative*, i.e., researchers are less likely to find significant differences that are not “real.” We chose to use this test along with the original bootstrap test because of this advantage, and also because the two tests are similar in spirit in that they rely on modern computational power instead of making many statistical assumptions. (Smucker, Allan and Carterette [25] recommend the randomisation test for *pairwise* significance testing.)

```

foreach pair of runs  $(r_1, r_2)$  do  $count(r_1, r_2) = 0$ ;
for  $b = 1$  to  $B$  do {
  create matrix  $\mathbf{X}^{*b}$  whose row  $t$  is a permutation of row  $t$  of  $\mathbf{X}$ 
    for every  $t \in T$ ;
   $max^{*b} = \max_i \bar{x}_i^{*b}$ ;  $min^{*b} = \min_i \bar{x}_i^{*b}$  where
     $\bar{x}_i^{*b}$  is the mean of  $i$ -th column vector of  $\mathbf{X}^{*b}$ ;
  foreach pair of runs  $(r_1, r_2)$  {
    if  $(max^{*b} - min^{*b} > |\bar{x}(r_1) - \bar{x}(r_2)|)$  where
       $\bar{x}(r_i)$  is the mean of the column vector for run  $r_i$  in  $\mathbf{X}$ 
       $count(r_1, r_2)++$ ;
  }
  foreach pair of runs  $(r_1, r_2)$  do  $ASL(r_1, r_2) = count(r_1, r_2)/B$ ;

```

Figure 4: Algorithm for obtaining the Achieved Significance Level with the two-sided, randomised Tukey’s HSD given a performance value matrix \mathbf{X} whose rows represent topics and columns represent runs [4].

```

foreach pair of runs  $(r_1, r_2)$  with a significant difference at  $\alpha$  do
   $\Delta_\alpha(r_1, r_2) = |mean(r_1) - mean(r_2)|$ ;
 $\Delta_\alpha = \min_{i,j} \Delta_\alpha(r_i, r_j)$ ;

```

Figure 5: Algorithm for estimating the performance Δ required for obtaining a significant difference at α with the randomised Tukey’s HSD test.

Let $t_{T,i}$ denote the i -th topic from a topic set T of size N , and let $M(t, r_j)$ denote the value of a metric M for a topic t and a run r_j . A paired bootstrap test for a given run pair (r_1, r_2) can be performed as shown in Figure 2: first, a vector \mathbf{z} of per-topic performances differences are obtained, and we set up a null hypothesis (H_0) saying that these values were sampled from a distribution whose population mean is zero; then, to construct an empirical distribution that obeys H_0 , a shifted vector \mathbf{w} is prepared and B bootstrap samples are obtained from it; then, for every trial b , the studentised statistic of \mathbf{z} (i.e., $t(\mathbf{z})$) is compared with the corresponding statistic for the bootstrap sample $(t(\mathbf{w}^{*b}))$; in this way, we obtain the *Achieved Significance Level* (ASL; a.k.a. p -value), which represents how likely \mathbf{z} would be under H_0 . As in any other significance testing, H_0 is rejected if $ASL < \alpha$.

Based on the bootstrap test, Sakai [18] also showed how to estimate the performance Δ required in order to achieve statistical significance at α given the topic set size N : the algorithm is shown in Figure 3. For example, if we have $B = 1,000$ bootstrap samples and $\alpha = 0.05$, we find the 50-th largest $|t(\mathbf{w}^{*b})|$ and record the corresponding non-studentised mean $|\bar{w}^{*b}|$ for every run pair. These values represent the borderline Δ ’s between significance and nonsignificance. Finally, to be conservative, we take the maximum value observed across all run pairs.

In contrast to pairwise tests such as the bootstrap test, the main idea behind Tukey’s HSD is that if the largest mean difference observed is not significant, then none of the other differences should be significant either. Given a set of runs, the null hypothesis is that there is no difference between *any* of the systems. Following Carterette [4], we perform randomised Tukey’s HSD as shown in Figure 4: from a given matrix \mathbf{X} whose element at (row i , column j) represents the performance of the j -th run for the i -th topic, we create B new matrices \mathbf{X}^{*b} by permutating each row at random; then, for every run pair, we compare the performance Δ of this run pair with the largest performance Δ observed within \mathbf{X}^{*b} . Finally, the ASL value is computed in a way similar to Figure 2, but for each run pair.

```

Disagreements = 0; Correct1 = 0; Correct2 = 0;
foreach pair of runs (r1, r2) do
  foreach topic t do {
    ΔM1 = M1(t, r1) - M1(t, r2);
    ΔM2 = M2(t, r1) - M2(t, r2);
    ΔMGS = MGS(t, r1) - MGS(t, r2);
    if( ΔM1 × ΔM2 < 0 ) // M1 and M2 disagree
      Disagreements ++;
    if( ΔM1 × ΔMGS ≥ 0 ) // M1 and MGS agree
      Correct1 ++;
    if( ΔM2 × ΔMGS ≥ 0 ) // M2 and MGS agree
      Correct2 ++;
  }
}
Intuitiveness(M1|M2, MGS) = Correct1/Disagreements;
Intuitiveness(M2|M1, MGS) = Correct2/Disagreements;

```

Figure 6: Algorithm for computing the intuitiveness of metrics M_1 and M_2 based on preference agreement with M^{GS} .

Using the results of the randomised Tukey’s HSD tests, we also try to estimate the performance Δ required to achieve a statistical significance at α for a given topic set size as shown in Figure 5: we simply take the smallest observed Δ from all the run pairs that were found to be significantly different.

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 (e.g., [22]). 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. Moreover, note that, by means of discriminative power, we are measuring the robustness of metrics to variations in the choice of topics and therefore the reliability of experiments: we are *not* discussing which particular differences are actually perceptible to the user. We do believe, however, that significance testing is one useful tool for making “real” improvements that may eventually add up to produce user-perceptible differences.

4.2 Intuitiveness Test

Sakai and Song [21] manually examined the actual ranked lists of documents to compare the intuitiveness of different diversity metrics, but this paper proposes a simple method for quantifying the intuitiveness. Suppose we want to compare two diversity metrics M_1 and M_2 . We deliberately choose a simple *Gold Standard metric* M^{GS} that should represent the intuitiveness, or the most important property that the diversity metrics should satisfy. For the purpose of search result diversification, the two most important properties are *diversity* and *relevance*. In the present study, we use intent recall (I-rec at l) to represent diversity, and *effective precision* (Ef-P at l) to represent relevance. Here, Ef-P is the proportion of documents that are *effectively relevant* to at least one intent: for informational intents, “effectively relevant” just means relevant; for each navigational intent, it means that only the first relevant document is counted as relevant and other “redundant” relevant documents are ignored. For example, the Ef-P for the example shown in Figure 1 (Section 3) is $3/5 = 0.8$, as the document at rank 4 is treated as nonrelevant. Note that these two gold standards are *set retrieval* metrics based on *binary* relevance: as different diversity metrics employ different rank-based discounting and different ways to de-

fine graded relevance, the gold standards should be chosen so that they are as agnostic to these differences as possible.

Given M_1, M_2 and M^{GS} (i.e., either I-rec or Ef-P), we measure the relative intuitiveness of the two diversity metrics in terms of *preference agreement* with M^{GS} as shown in Figure 6. In this pseudocode, *Disagreement* is the number of ranked list pairs for which the two diversity metrics disagreed with each other as to which list is better; *Correct₁* is the number of ranked list pairs from the disagreements, for which M_1 agrees with the “correct judgment” of M^{GS} , and so on. In the pseudocode, note that if ΔM^{GS} is zero (i.e., the gold standard says that the two ranked lists are tied), this case is counted as a “correct” case. We found that ties actually occur quite often with “crude” metrics such as I-rec.

Note also that we focus on the disagreements between M_1 and M_2 rather than the entire set of ranked list pairs. (We have a total of 4,560 pairs: 24 topics \times 190 run pairs.) This is because we already know that different diversity metrics are generally highly correlated to one another [21]. Thus, Figure 6 enables us to discuss “which metric is more intuitive *than the other*” assuming that the gold standard truly represents intuitiveness.

We can expect metrics such as D $\#$ -measures, DIN $\#$ -measures and P+Q $\#$ to show good intuitiveness results when I-rec is used as the gold standard, since these metrics directly depend on I-rec by means of Eq. 7 and the like. Also, we can expect DIN $\#$ -measures and P+Q $\#$ to show good results when Ef-P is used as the gold standard, since these metrics all rely on the basic idea of ignoring redundant documents for navigational intents⁶. In short, it would not be surprising if our proposed metrics do well in our intuitiveness experiments. The contribution here, however, is that we are able to quantify exactly how much some of these metrics outperform the other metrics, including the popular α -nDCG.

The above method considers diversity (I-rec) and relevance (Ef-P) one at a time. However, what we really want are intuitive evaluation metrics that consider both. We therefore extend the algorithm shown in Figure 6 to handle two gold-standard metrics M_1^{GS} and M_2^{GS} (which in this paper are I-rec and Ef-P): in this case, *Correct₁* is incremented only if M_1 agrees with M_1^{GS} and with M_2^{GS} , and so on.

5. EXPERIMENTS

5.1 Data

For evaluating different diversity metrics in terms of discriminative power and intuitiveness, we used the graded-relevance version of the TREC 2009 Web track diversity test collection with *Category A* runs [6]. The original TREC data has binary per-intent relevance assessments, but this version contains *L3* (highly relevant), *L2* (relevant) and *L1* (partially relevant) documents for each intent, which were defined based on judgements from multiple assessors [21]⁷. From the official 50 topics, we selected those that had at least one navigational subtopic (i.e., intent), which resulted in 24 topics. Some statistics of this data set are shown in Table 1. As shown in the table, our data set contains 68 informational and 31 navigational intents, with a total of 2,635 relevant documents for the informational intents and 198 for the navigational intents; we use the uniform and nonuniform intent probabilities of Sakai and

⁶Note, however, that while both DIN $\#$ -measures and Ef-P takes the first relevant document for each navigational document as relevant, P+Q $\#$ goes down to the preferred rank rp as was discussed in Section 3.2.

⁷The graded relevance assessments are available from the author on request.

Table 1: Test collection statistics.

#Documents	Approx. one billion Web pages (ClueWeb09).
#Topics	24, with at least one navigational intent (17 faceted; 7 ambiguous).
#Intents	99, with at least one relevant document (68 informational; 31 navigational)
Mean and Range of #Intents/topic	4.1 [1, 6] (all); 2.8 [1, 5] (informational); 1.3 [1, 3] (navigational) across 24 topics.
Intent probabilities for n intents	Uniform: j -th intent has the probability $1/n$; Nonuniform: j -th intent has the probability $2^{n-j+1} / \sum_{k=1}^n 2^k$.
#Relevant	2,635 (1,465 $L3$ -relevant; 663 $L2$ -relevant; 507 $L1$ -relevant) across 99 intents; 2,437 (1,328 $L3$ -relevant; 620 $L2$ -relevant; 489 $L1$ -relevant) across 68 informational intents; 198 (137 $L3$ -relevant; 43 $L2$ -relevant; 18 $L1$ -relevant) across 31 navigational intents.
Mean and Range of #Intents/document	1.19 [1, 4] across 2,223 unique relevant summed across topics.
#Runs	20 Category A runs selected at random.

Song [21], and the 20 sampled runs from the same study, which gives us 190 run pairs.

Following previous work [17, 18, 20, 21], we used $B = 1,000$ for the bootstrap tests. On the other hand, as we had no previous experience in using the randomised Tukey’s HSD, we determined the value of B through a preliminary experiment: Smucker, Allan and Carterette [25] used $B = 100,000$ for their *pairwise* randomisation test but we thought that a fewer number of trials may suffice. Figure 7(b) shows the *ASL curves* [18] for $D\#$ -nDCG with the uniform intent probabilities based on the randomised Tukey’s HSD test for different values of B : the y -axis represents the ASL and the x axis represents the 190 run pairs sorted by the ASL. The graphs are somewhat cluttered but that is exactly the point: for example, the curve for $B = 5,000$ almost completely overlaps with that for $B = 10,000$. Based on these results, we use $B = 5,000$ for randomised Tukey’s HSD. For reference, Figure 7(a) shows a similar set of graphs for the bootstrap test: it can be observed that $B = 1,000$ is probably sufficient, and that much lower ASLs are obtained compared to Tukey’s HSD. This demonstrates the fact that pairwise tests may “find” significant differences that are not substantial. Note that the family-wise error rate (See Section 4.1) for any pairwise tests at $\alpha = 0.05$ with 190 run pairs is $1 - 0.95^{190} > 0.9999$.

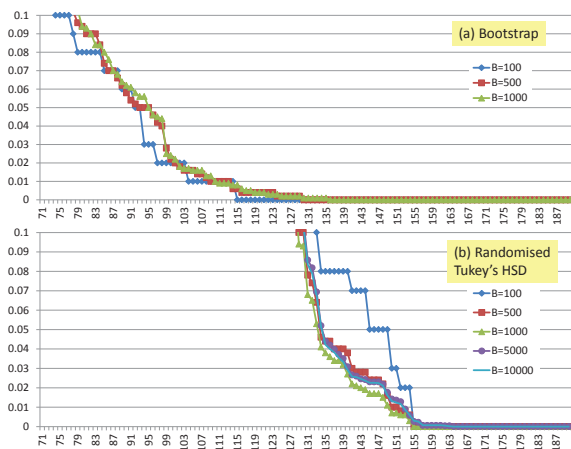


Figure 7: Effect of B on the accuracy of the ASL curve for $D\#$ -nDCG@10 (Uniform $Pr(i|q)$). y -axis: ASL (i.e., p -value); x -axis: run pairs sorted by ASL.

5.2 Evaluation Toolkit

For computing all evaluation metrics, we used a version of the *NTCIREVAL* toolkit [19]⁸. The only exception was α -nDCG: we

⁸Available at <http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html>.

used the official α -nDCG values from TREC (with $\alpha = 0.5$) as implementing this metric requires a greedy approximation of the ideal ranked list [8]. For all metrics, we used the document cutoff of $l = 10$ as we are interested in evaluating the first SERP, the entry-point page for different user intents.

5.3 Discriminative Power Results

Figures 8 and 9 show the ASL curves of some selected diversity metrics, based on the bootstrap test and the randomised Tukey’s HSD, respectively. Parts (a) of these figures show the results with the uniform intent probability distribution: α -nDCG and I-rec are included here as these two metrics do not utilise intent probabilities. Parts (b) of these figures show the results with the nonuniform distribution: D-nDCG, DIN-nDCG and P+Q are included here to highlight the effect of combining these metrics with I-rec and thereby obtaining $D\#$ -nDCG, $DIN\#$ -nDCG and $P+Q\#$. We want metrics that are discriminative, i.e., those that are closer to the origin in the figures.

Tables 2 and 3 cut Figures 8 and 9 in half at $\alpha = 0.05$ to quantify discriminative power and the performance Δ required for achieving statistical significance with 24 topics. For example, Table 2(a) shows that the discriminative power of I-rec according to the bootstrap test at $\alpha = 0.05$ is $(100/190) = 52.6\%$ (i.e., 100 significantly different run pairs were found) and the Δ required for achieving statistical significance is around 0.20.

First, by comparing the bootstrap and the randomised Tukey’s HSD results (i.e., Figure 8 vs. Figure 9 and Table 2 vs. Table 3), it can be observed that:

- The relative performances of the different metrics are generally similar with these two tests, although it is not clear why P+Q does relatively well with the randomised Tukey’s HSD (Figure 9(b)) but not with the bootstrap test (Figure 8(b)).
- The randomised Tukey’s HSD is substantially more conservative than the bootstrap test, as it is clear from the contrast between Figures 8 and 9. For example, at $\alpha = 0.05$, the discriminative power of I-rec according to the bootstrap is 52.6% (Table 2(a)), while that according to the randomised Tukey’s HSD is only 26.8% (Table 3(a)): that is, about half of the significant differences obtained with the bootstrap test are *not* significant with the randomised Tukey’s HSD. (This set of significant differences obtained by the randomised Tukey’s HSD is a true subset of the set of significant differences obtained by the bootstrap test.)
- The performance Δ ’s as estimated with the randomised Tukey’s HSD are similar to the corresponding values based on the bootstrap test. For example, with the uniform setting, the performance Δ required for achieving a statistical significance with P+Q# given 24 topics is 0.15 according to both tests (Tables 2(a) and 3(a)).

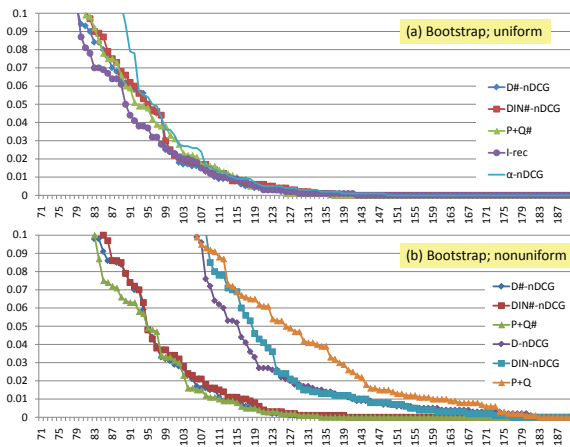


Figure 8: ASL curves based on the bootstrap test. y -axis: ASL (i.e., p -value); x -axis: run pairs sorted by ASL.

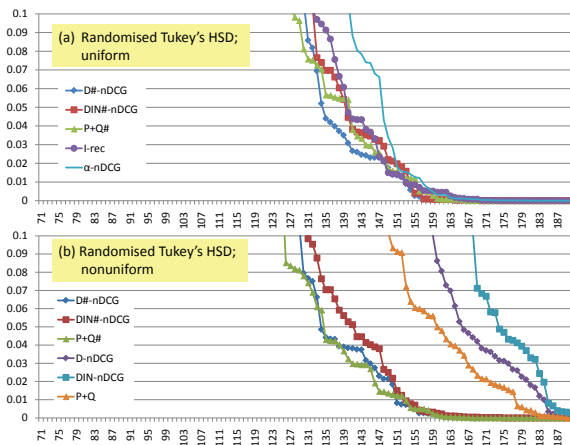


Figure 9: ASL curves based on the randomised Tukey's HSD. y -axis: ASL (i.e., p -value); x -axis: run pairs sorted by ASL.

The above observations suggest that the randomised Tukey's HSD is a good alternative to the pairwise bootstrap test for the purpose of comparing evaluation metrics. Also, given a set of available runs, researchers are encouraged to make use of all of these runs in significance testing, as focussing on a particular set of runs (by means of a pairwise test) may often lead to wrong conclusions [4].

Next, by comparing the different metrics in terms of discriminative power as shown in Figures 8 and 9 and Tables 2 and 3, it can be observed that:

- $D_{\#}$ -nDCG and $P+Q_{\#}$ are comparable to I-rec and $D_{\#}$ -nDCG in terms of discriminative power (Figures 8 and 9)⁹.
- α -nDCG may be slightly less discriminative than the above best metrics (Figures 8(a) and 9(a)). This difference is evident particularly when α is large (e.g., $\alpha \geq 0.1$). (Again, this α is the Type I Error probability, not the redundancy parameter of α -nDCG.)
- Combination with I-rec dramatically boosts the discriminative power of all diversity metrics (e.g., compare $P+Q_{\#}$ with $P+Q$ in Figures 8(b) and 9(b));

⁹ $D_{\#}$ -nDCG will never *outperform* $D_{\#}$ -nDCG in terms of discriminative power, because it is based on fewer data points (i.e., documents treated as relevant) than $D_{\#}$ -nDCG when not identical to it.

Table 2: Discriminative power / performance Δ of diversity metrics based on the bootstrap test at $\alpha = 0.05$.

(a) uniform			(b) nonuniform		
I-rec	52.6%	0.20	$P+Q_{\#}$	50.5%	0.16
$P+Q_{\#}$	51.6%	0.15	$D_{\#}$ -nDCG	50.5%	0.14
$D_{\#}$ -nDCG	50.0%	0.16	DIN $_{\#}$ -nDCG	50.5%	0.16
DIN $_{\#}$ -nDCG	50.0%	0.15	$D_{\#}$ -Q	48.9%	0.18
$D_{\#}$ -Q	50.0%	0.14	DIN $_{\#}$ -Q	48.9%	0.16
DIN $_{\#}$ -Q	49.5%	0.16	D-nDCG	39.5%	0.12
α -nDCG	49.5%	0.15	DIN-nDCG	37.9%	0.12
D-nDCG	43.2%	0.14	D-Q	35.3%	0.12
DIN-nDCG	41.6%	0.14	$P+Q$	33.7%	0.15
D-Q	36.8%	0.12	DIN-Q	33.2%	0.14
DIN-Q	34.7%	0.15			
$P+Q$	34.2%	0.12			

Table 3: Discriminative power / performance Δ of diversity metrics based on the randomised Tukey's HSD test at $\alpha = 0.05$.

(a) uniform			(b) nonuniform		
$D_{\#}$ -nDCG	29.5%	0.17	$D_{\#}$ -nDCG	30.0%	0.17
$D_{\#}$ -Q	26.8%	0.16	$P+Q_{\#}$	29.5%	0.16
DIN $_{\#}$ -nDCG	26.8%	0.17	$D_{\#}$ -Q	26.8%	0.15
I-rec	26.8%	0.23	DIN $_{\#}$ -nDCG	25.8%	0.16
$P+Q_{\#}$	26.3%	0.15	DIN $_{\#}$ -Q	22.6%	0.15
DIN $_{\#}$ -Q	23.7%	0.15	$P+Q$	15.8%	0.12
α -nDCG	22.6%	0.17	D-nDCG	13.2%	0.15
D-nDCG	18.9%	0.14	DIN-nDCG	8.9%	0.13
$P+Q$	18.4%	0.09	D-Q	2.6%	0.13
DIN-nDCG	15.8%	0.13	DIN-Q	0.5%	0.12
D-Q	6.3%	0.12			
DIN-Q	3.7%	0.11			

Table 4: Comparison of significantly different run pairs (randomised Tukey's HSD at $\alpha = 0.05$; uniform setting).

	α -nDCG	$D_{\#}$ -nDCG	DIN $_{\#}$ -nDCG	$P+Q_{\#}$
I-rec	13/38/5	0/51/5	1/50/1	3/48/2
α -nDCG	-	1/42/14	4/39/12	4/39/11
$D_{\#}$ -nDCG	-	-	5/51/0	6/50/0
DIN $_{\#}$ -nDCG	-	-	-	3/48/2

- D-Q and DIN-Q are the least discriminative metrics among the ones we examined (e.g., see bottom of Table 3(b)). Moreover, in the tables, DIN $_{\#}$ -Q is never more discriminative than DIN $_{\#}$ -nDCG, and $D_{\#}$ -Q is never more discriminative than $D_{\#}$ -nDCG. (For these reasons, DIN($\#$)-Q and D($\#$)-Q are not shown in the two figures.)

The above observations suggest that DIN $_{\#}$ -nDCG and $P+Q_{\#}$ are promising as metrics that explicitly takes into account whether each intent is informational or navigational. The high discriminative power comes mostly from the simple I-rec metric. Note, however, that these results only suggest that DIN $_{\#}$ -nDCG and $P+Q_{\#}$ are statistically reliable and consistent: they say nothing about whether they are right or wrong. Hence we discuss the intuitiveness of these metrics in Section 5.4. Based on the above results, we hereafter focus our attention to DIN $_{\#}$ -nDCG and $P+Q_{\#}$ as well as $D_{\#}$ -nDCG and α -nDCG for comparison purposes.

Table 4 provides a further analysis of some of the results from Table 3(a), i.e., the randomised Tukey's HSD results with the uniform setting. The table shows the degree of overlap between the sets of significantly different pairs for I-rec, α -nDCG, $D_{\#}$ -nDCG, DIN $_{\#}$ -nDCG and $P+Q_{\#}$. For example, it can be observed from the rightmost column that I-rec and $P+Q_{\#}$ have 48 run pairs in common, and that these two metrics obtained $3 + 48 = 51$ significant differences and $48 + 2 = 50$ significant differences, respectively. (These correspond to the discriminative power values of $51/190 = 26.8\%$ and $50/190 = 26.3\%$ in Table 3(a).) The main message this table

Table 5: Kendall’s τ / Symmetric τ_{ap} for ranking the 20 runs (uniform setting). Values higher than 0.9 are shown in bold for convenience.

	α -nDCG	$D_{\#}$ -nDCG	$DIN_{\#}$ -nDCG	$P+Q_{\#}$
I-rec	.74/.80	.91/.93	.92/.94	.94/.93
α -nDCG	-	.79/.83	.80/.94	.78/.84
$D_{\#}$ -nDCG	-	-	.99/.99	.95/.95
$DIN_{\#}$ -nDCG	-	-	-	1/1

Table 6: Intuitiveness based on preference agreement. For each metric pair, the higher score is shown in bold. Disagreements are shown in parentheses.

(a) Baseline: I-rec (“diversity”)			
	$D_{\#}$ -nDCG	$DIN_{\#}$ -nDCG	$P+Q_{\#}$
α -nDCG	.597/ .995 (236)	.607/ .996 (242)	0.573/ 1 (246)
$D_{\#}$ -nDCG	-	1/1 (19)	.908/ 1 (120)
$DIN_{\#}$ -nDCG	-	-	.907/ 1 (118)
(b) Baseline: Effective Precision (“relevance”)			
	$D_{\#}$ -nDCG	$DIN_{\#}$ -nDCG	$P+Q_{\#}$
α -nDCG	.623/ .733 (236)	.616/ .748 (242)	.646/ .654 (246)
$D_{\#}$ -nDCG	-	.474/ .842 (19)	.800 /.625 (120)
$DIN_{\#}$ -nDCG	-	-	.831 /.593 (118)
(c) Baseline: I-rec AND Effective Precision			
	$D_{\#}$ -nDCG	$DIN_{\#}$ -nDCG	$P+Q_{\#}$
α -nDCG	.398/ .729 (236)	.397/ .744 (242)	.390/ .654 (246)
$D_{\#}$ -nDCG	-	.474/ .842 (19)	.708 /.625 (120)
$DIN_{\#}$ -nDCG	-	-	.737 /.593 (118)

conveys is that these metrics are quite similar to each other *when averaged across topics*.

Table 5 shows the Kendall’s τ and (the symmetric version of) τ_{ap} proposed by Yilmaz, Aslam and Robertson [27] for ranking the 20 runs by the aforementioned five metrics. τ_{ap} compares the similarity of two run rankings based on pairwise swaps just like τ , but is more “top-heavy.” It can be observed that the rankings by I-rec, $D_{\#}$ -nDCG, $DIN_{\#}$ -nDCG and $P+Q_{\#}$ all resemble each other, quite naturally as the “ $\#$ ” represents linear combination with I-rec. Perhaps what is more interesting is that the ranking by $DIN_{\#}$ -nDCG and $P+Q_{\#}$ are actually identical as indicated by the τ and τ_{ap} values of 1, despite the different approaches they employ (See Section 3).

Tables 4 and 5 show how similar the five diversity metrics are *on average*; below we focus on individual cases where they differ.

5.4 Intuitiveness Results

Table 6 show the intuitiveness scores for α -nDCG, $D_{\#}$ -nDCG, $DIN_{\#}$ -nDCG and $P+Q_{\#}$ computed using the preference agreement algorithm shown in Figure 6: Part (a) uses I-rec as the gold-standard and therefore represents how the diversity metrics favour diversified results like they should; Part (b) uses Ef-P as the gold-standard and therefore represents how they favour the result with more relevant documents like they should (while ignoring redundant relevant documents for navigational intents). Part (c) computes the intuitiveness scores by requiring that the diversity metrics agree with both I-rec and Ef-P. For example, Table 6(a) shows that, if we compare α -nDCG and $D_{\#}$ -nDCG in terms of diversity, there are 236 disagreements, and that while the intuitive score for α -nDCG

is only .597, that for $D_{\#}$ -nDCG is .995. This means that, given a pair of ranked lists for which α -nDCG and $D_{\#}$ -nDCG disagree with each other, $D_{\#}$ -nDCG is far more likely to agree with I-rec on the preference than α -nDCG. The relative results can be summarised as follows:

- In terms of diversity (Part (a)), $D_{\#}$ -nDCG, $DIN_{\#}$ -nDCG and $P+Q_{\#}$ outperform α -nDCG; $P+Q_{\#}$ outperforms $D_{\#}$ -nDCG and $DIN_{\#}$ -nDCG; and therefore $P+Q_{\#}$ is the winner (note that $P+Q_{\#}$ agrees 100% with I-rec in the rightmost column of Table 6(a)).
- In terms of relevance (Part (b)), $D_{\#}$ -nDCG, $DIN_{\#}$ -nDCG and $P+Q_{\#}$ outperform α -nDCG; $D_{\#}$ -nDCG outperforms $P+Q_{\#}$; $DIN_{\#}$ -nDCG outperforms $D_{\#}$ -nDCG and $P+Q_{\#}$; and therefore $DIN_{\#}$ -nDCG is the winner.
- In terms of both diversity and relevance (Part (c)), $D_{\#}$ -nDCG, $DIN_{\#}$ -nDCG and $P+Q_{\#}$ outperform α -nDCG; $D_{\#}$ -nDCG outperforms $P+Q_{\#}$; $DIN_{\#}$ -nDCG outperforms $D_{\#}$ -nDCG and $P+Q_{\#}$; and therefore $DIN_{\#}$ -nDCG is the winner.

Recall that these results should be regarded with a grain of salt: first, it is not surprising that $D_{\#}$ -nDCG, $DIN_{\#}$ -nDCG and $P+Q_{\#}$ behave similarly to I-rec (See Eq. 7); second, it is not surprising that $DIN_{\#}$ -nDCG behaves similarly to Ef-P, since they both look at the first retrieved relevant document for every navigational intent. However, note that our results say much more: they suggest that $P+Q_{\#}$ may be the most intuitive “diversity-oriented” diversity metric, and that $DIN_{\#}$ -nDCG may be the most intuitive “relevance-oriented” diversity metric and the best metric that takes both diversity and relevance into account. Moreover, they quantitatively show the advantages of these metrics over α -nDCG.

Note also that our simple approach to discussing intuitiveness based on preference agreement is applicable to *any* pair of evaluation metrics provided that an appropriate gold-standard metric can be defined.

6. CONCLUSIONS

In this study, we proposed new evaluation metrics called $DIN_{\#}$ -measures and $P+Q_{\#}$ which incorporate the explicit knowledge of informational and navigational intents into diversity evaluation. Like Intent-Aware metrics and $D_{\#}$ -measures, these metrics can handle intent probabilities and per-intent graded relevance. (Recall that α -nDCG can handle neither.) We also proposed a simple method for comparing the intuitiveness of a given pair of metrics quantitatively. Our main experimental findings are:

- In terms of discriminative power, the proposed metrics, $DIN_{\#}$ -nDCG and $P+Q_{\#}$, are comparable to intent recall and $D_{\#}$ -nDCG, and possibly superior to α -nDCG;
- In terms of preference agreement with intent recall, $P+Q_{\#}$ is superior to other diversity metrics and therefore may be the most intuitive as a metric that emphasises diversity; and
- In terms of preference agreement with effective precision, $DIN_{\#}$ -nDCG is superior to other diversity metrics and therefore may be the most intuitive as a metric that emphasises relevance. Moreover, $DIN_{\#}$ -nDCG may be the most intuitive as a metric that considers both diversity and relevance.

In addition, we demonstrated that the randomised Tukey’s Honestly Significant Differences test (described by Carterette [4]) that takes the entire set of available runs into account is substantially more conservative than the paired bootstrap test that only considers one

run pair at a time. We therefore recommend the former approach for significance testing when a set of runs is available for evaluation.

Finally, limitations to the present study include the following:

1. As was discussed in Section 3, $DIN_{\#}$ -measures and $P+Q_{\#}$ are *imperfect* metrics, in that they do not range fully between 0 and 1. However, we regard this as a cost of taking into account the distinction between informational and navigational intents and yet keeping the metrics simple to undertake and to compute.
2. Our experiments do not involve human participants: we believe that our approach and user-based studies such as the work by Sanderson *et al.* [23] are complementary. Note that it is not straightforward to conduct a user study for diversity metrics, as a diversified SERP is intended for a population of users sharing the same query but having different intents, as opposed to a small group of participants.
3. Our experiments rely on a single test collection, with only 24 topics and *artificial* intent probabilities [21]. We plan to expand our experiments by adding graded relevance assessments to the latest TREC web diversity test collections and/or utilising the NTCIR INTENT test collections which come with intent probabilities obtained through assessor voting [26].
4. While the proposed metrics leverage the explicit knowledge of whether each intent is informational or navigational, there is another aspect that is available in the TREC diversity test collections which we did not consider, namely, the distinction between *ambiguous* and *faceted* topics [6]. Clarke, Kolla and Vechtomova [9] have briefly discussed this in the context of extending α -nDCG. However, the challenge here would be how to keep the evaluation metric simple and intuitive.

7. 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] C. Brandt, T. Joachims, Y. Yue, and J. Bank. Dynamic ranked retrieval. In *Proceedings of ACM WSDM 2011*, 2011.
- [3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2), 2002.
- [4] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1), 2012.
- [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*, 2009.
- [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.
- [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 *Proceedings of ICTIR 2009*, 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*, pages 475–484, 2011.
- [11] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44:1251–1266, 2008.
- [12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
- [13] T. Leelanupab, G. Zuccon, and J. M. Jose. A query-basis approach to parametrizing novelty-biased cumulative gain. In *Proceedings of ICTIR 2011*, pages 327–331, 2011.
- [14] C. Macdonald, J. Wang, and C. Clarke, editors. *Proceedings of the Diversity in Document Retrieval 2001 Workshop*. 2011.
- [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, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *Proceedings of ACM SIGIR 2010*, pages 603–610, 2010.
- [17] T. Sakai. Bootstrap-based comparisons of IR metrics for finding one relevant document. In *Proceedings of AIRS 2006 (LNCS 4182)*, pages 374–389, 2006.
- [18] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532, 2006.
- [19] T. Sakai. NTCIREVAL: A generic toolkit for information access evaluation. In *Proceedings of FIT 2011 (Volume 2)*, pages 23–30, 2011.
- [20] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA 2008*, pages 30–41, 2008.
- [21] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of ACM SIGIR 2011*, pages 1043–1042, 2011.
- [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. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of ACM SIGIR 2011*, pages 595–604, 2011.
- [25] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance test for information retrieval evaluation. In *Proceedings of ACM CIKM 2007*, pages 623–632, 2007.
- [26] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT Task. In *Proceedings of NTCIR-9*, pages 82–105, 2011.
- [27] 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.
- [28] 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.