Journal of Information Processing Vol.21 No.1 145–155 (Jan. 2013)

[DOI: 10.2197/ipsjjip.21.145]

Regular Paper

Web Search Evaluation with Informational and Navigational Intents

Tetsuya Sakai[1,a)]

Received: May 23, 2012, Accepted: October 10, 2012

Abstract: Given an ambiguous or underspecified web search 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 the condordance test 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, DIN♯-nDCG and P+Q♯, are comparable to intent recall and D♯-nDCG, and possibly superior to α-nDCG; (b) In terms of the *concordance test* which quantifies the agreement of a diversity metric with a gold standard metric that represents a basic desirable property, DIN♯-nDCG is superior to other diversity metrics in its ability to reward both diversity and relevance at the same time. Moreover, both D♯-nDCG and DIN♯-nDCG significantly outperform α-nDCG in their ability to reward diversity, to reward relevance, and to reward both at the same time. 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.

Keywords: diversified search, evaluation, intents, metrics, search result diversification, web search

1. Introduction

Web 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., Refs. [1], [10], [14], [15], [27]). TREC (Text Retrieval Conference)[*1] began a diversity task in the Web track in 2009 [6][*2], and NTCIR (NII Testbeds and Community for Information access Research)[*3] concluded its first INTENT task for mining intents and selectively diversifying search results in 2011 [29][*4]. These tasks evaluate plain, ranked lists of diversified web pages, 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 [23] conducted an extensive study of different diversity metrics in terms of *discriminative power* [17], [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.) Whereas, Sakai and Song discussed intuitiveness by manually examining pairs of ranked lists, and showed that a family of metrics called *D♯-measures* [23] 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.

While the three problems mentioned above do not apply to D♯-measures, the manual analysis by Sakai and Song [23] suggested

[1] Microsoft Research Asia, Beijing 100080, P.R.China
[a)] tetsuyasakai@acm.org

[*1] http://trec.nist.gov/
[*2] http://plg.uwaterloo.ca/ trecweb/
[*3] http://research.nii.ac.jp/ntcir/
[*4] http://research.microsoft.com/en-us/people/tesakai/intent2.aspx

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[*5]. We propose new diversity evaluation metrics called *DIN♯-measures* and *P+Q♯*, as well as the *concordance test* for comparing the intuitiveness of a given pair of metrics quantitatively. Our main experimental findings are:

(a) In terms of *discriminative power* [17], [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;

(b) In terms of the *concordance test* (introduced in Section 4.2) which quantifies the agreement of a diversity metric with a gold standard metric that represents a basic desirable property, DIN♯-nDCG is superior to other diversity metrics in its ability to reward both diversity and relevance at the same time. Moreover, both D♯-nDCG and DIN♯-nDCG significantly outperform α-nDCG in their ability to reward diversity, to reward relevance, and to reward both at the same time.

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* [17], [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 the concordance test. Section 5 describes our experiments and reports on discriminative power and concordance test results. Finally, Section 6 concludes this paper.

[*5] This paper is a revised version of an international conference paper published in April 2012 [21].

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 β (≥ 0) is a user persistence parameter which is set to 1 throughout this study. Then *Q-measure* [17], [18], [22] is defined as:

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

where R is the total number of known relevant documents and L is the size of the ranked list. Note that $\beta = 0$ reduces Q-measure to the well-known 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* [5] and P^+ [16], [19] 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 cutoff[*6].

Formally, P+ is defined as [16], [19]:

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

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

Sakai [16], [19] 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 relevant documents that are *treated* as relevant. Similarly, as was mentioned earlier, Sakai and Song [23] 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 [22]. 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 [22] 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 [6]). Formally, let $J_n(r) = 1$ if a document at rank r is relevant to the n-th nugget and 0 otherwise; let $C_n(r) = \sum_{k=1}^{r} J_n(r)$, 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

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

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).

α-nDCG and IA metrics are not guaranteed to range between 0 and 1: in the case of α-nDCG, computing an ideal list based on nuggets is NP-complete; in the case of IA metrics, 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 [23] proposed to plot D-measures against *intent recall* (a.k.a. *I-rec* or *subtopic recall* [31], 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:

$$D\sharp\text{-}measure = \gamma I\text{-}rec + (1-\gamma)D\text{-}measure \quad (7)$$

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 [23]. The NTCIR-9 INTENT task also used D♯-nDCG with $\gamma = 0.5$ as the primary metric for ranking participating systems [29].

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 diversity test collection [6].

Sakai and Song [23] 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

[*6] Suppose 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.

assessments added to the TREC 2009 Web diversity test 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 the concordance test for quantifying the relative intuitiveness of diversity metrics in this present study. In a sequel to Sakai and Song [23], Sakai and Song [24] showed that D(♯)-measures are superior to ERR-IA in terms of discriminative power *and* the concordance test using the NTCIR-9 INTENT test collections as well as the TREC 2009 Web diversity test collection.

Using the Amazon Mechanical Turk framework and the TREC 2009 Web diversity test collection with *binary* relevance assessments, Sanderson et al. [26] 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 concordance test 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 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, the DIN-measure family[*7], is identical to the D-measure family 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 **Fig. 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.

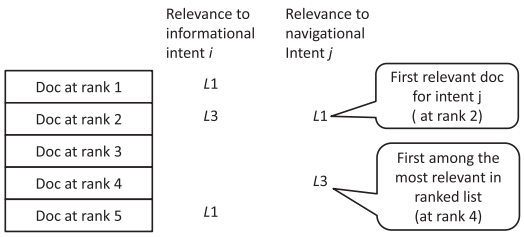

Fig. 1 An example ranked list for a query with one informational intent and one navigational intent.

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:

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

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

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

where

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

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 to boost diversity relative to relevance (Recall Eq. (7)). We call the resultant metrics DIN♯-metrics. In this paper, 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

[*7] DIN stands for: Diversification for Informational and Navigational intents.

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 Fig. 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 Fig. 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 intent recall to emphasise diversity in a way similar to Eq. (7), and call the resultant metric P+Q♯. Note that Sakai and Song [23], [24] did not consider the combination of IA metrics with I-rec, although it is also possible[*8].

4. Evaluating Evaluation Metrics

This section describes two methods for comparing the "goodness" of diversity metrics: *discriminative power* [17], [18], which represents the statistical reliability of a metric, and the *concordance 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 [17]. 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 $1 - (1 - \alpha)^k$: 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 [4]. 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 [28] have recommended the randomisation test for *pairwise* significance testing.)

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 **Fig. 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 [17] also showed how to estimate the performance delta (Δ) required in order to achieve statistical significance at α given the topic set size N: the algorithm is shown in **Fig. 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 $|\overline{w}^{*b}|$ for ev-

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

Fig. 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 [17].

[*8] Combining 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.

ery 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 **Fig. 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 Fig. 2, but for each run pair.

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 **Fig. 5**: we simply take the smallest observed Δ from all the run pairs that were found to be significantly different.

For more details on the bootstrap and the randomised version of Tukey's HSD test, we refer the reader to Sakai [17] and Carterette [4], respectively. Note that this paper does not propose any new statistical significant tests.

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., Ref. [25]). 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.

```
foreach pair of runs (r1, r2) do
    if( |t(w)^{*b'}| is the Bα-th largest value in {|t(w)^{*b}|})
        Δα(r1, r2) = |w̄^{*b'}|;
Δα = max_{i,j} Δα(ri, rj);
```

Fig. 3 Algorithm for estimating the performance Δ required for obtaining a significant difference at α with the paired bootstrap test [17].

```
foreach pair of runs (r1, r2) do count(r1, r2) = 0;
for b = 1 to B do {
    create matrix X^{*b} whose row t is a permutation of row t of X
                         for every t ∈ T;
    max^{*b} = max_i x̄_i^{*b}; min^{*b} = min_i x̄_i^{*b} where
          x̄_i^{*b} is the mean of i-th column vector of X^{*b};
foreach pair of runs (r1, r2) {
    if( max^{*b} − min^{*b} > |x̄(r1) − x̄(r2)| where
        x̄(ri) is the mean of the column vector for run ri in X )
        count(r1, r2) + +;
}
foreach pair of runs (r1, r2) do ASL(r1, r2) = count(r1, r2)/B;
```

Fig. 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 (r1, r2) with a significant difference at α do
    Δα(r1, r2) = |mean(r1) − mean(r2)|;
Δα = min_{i,j} Δα(r1, r2);
```

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

4.2 Concordance Test

Sakai and Song [23] manually examined the actual ranked lists of documents to compare the intuitiveness of different diversity metrics, but here we propose the concordance test for quantifying the intuitiveness. Suppose we want to compare two diversity metrics M_1 and M_2. We choose a deliberately simple *Gold Standard metric* M^{GS} that should represent the intuitiveness, i.e., 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 nagivational 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 Fig. 1 (Section 3) is $3/5 = 0.6$, as the document at rank 4 is treated as nonrelevant. Note that the gold standards themselves are not good enough as stand-alone diversity metrics: they ignore document ranks, graded relevance, and intent probabilities. We use them to separate out and test a particular property of a more complex metric.

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 concordance with M^{GS} as shown in **Fig. 6**. In this pseudocode,

```
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);
        ΔM^GS = M^GS(t, r1) − M^GS(t, r2);
        if( ΔM1 × ΔM2 < 0 ){ // M1 and M2 disagree
            Disagreements + +;
            if( ΔM1 × ΔM^GS ≥ 0) ) // M1 and M_GS agree
                Correct1 + +;
            if( ΔM2 × ΔM^GS ≥ 0) ) // M2 and M_GS agree
                Correct2 + +;
        }
    }
Concordance(M1|M2, M^GS) = Correct1/Disagreements;
Concordance(M2|M1, M^GS) = Correct2/Disagreements;
```

Fig. 6 Concordance test algorithm for metrics M_1 and M_2 based on preference agreement with M^{GS}.

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 across 99 intents (1,465 $L3$-relevant; 663 $L2$-relevant; 507 $L1$-relevant); 2,437 across 68 informational intents (1,328 $L3$-relevant; 620 $L2$-relevant; 489 $L1$-relevant); 198 across 31 navigational intents (137 $L3$-relevant; 43 $L2$-relevant; 18 $L1$-relevant).
Mean and Range of #Intents/document	1.19 [1, 4] for 2,223 unique relevant documents across topics.
#Runs	20 Category A runs selected at random.

Disagreement is the number of ranked list pairs for which the two diversity metrics disagreed with each other as to which list is better; $Correct_1$ 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 × 190 run pairs.) This is because we already know that different diversity metrics are generally highly correlated to one another [23]. Thus, Fig. 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 concordance test 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 intents[*9]. In short, it would not be surprising if our proposed metrics do well in our concordance experiments. The contribution here, however, is that we are able to quantify exactly how often some of these metrics outperform the other metrics, including the popular α-nDCG.

The above method considers diversity (I-rec) and effective 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 Fig. 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, $Correct_1$ 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 the concordance test, we used the *graded relevance* version of the TREC 2009 Web diversity test collection with *Category A* runs [6], which we obtained from Sakai and Song [23]. 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. 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 nagivational 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 Song [23], and the 20 sampled runs from the same study, which gives us 190 run pairs.

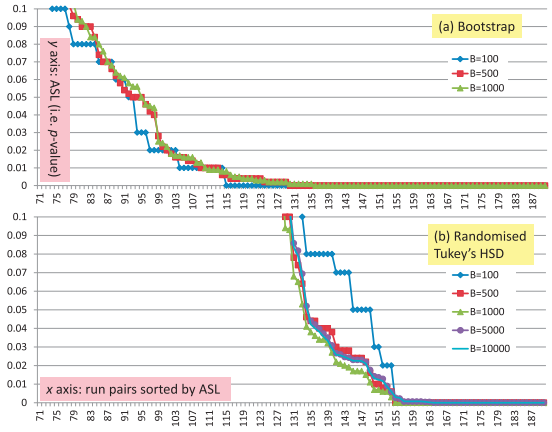

Fig. 7 Effect of B on the accuracy of the ASL curve for D♯-nDCG@10 (Uniform $Pr(i|q)$).

Following previous work [16], [17], [22], [23], 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 [28] 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* [17] 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, Fig. 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 α = 0.05 with 190 run pairs is $1 - 0.95^{190} > 0.9999$.

[*9] Note, 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.

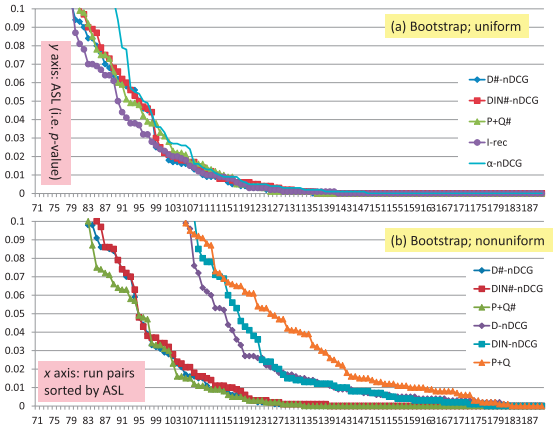

Fig. 8 ASL curves based on the bootstrap test.

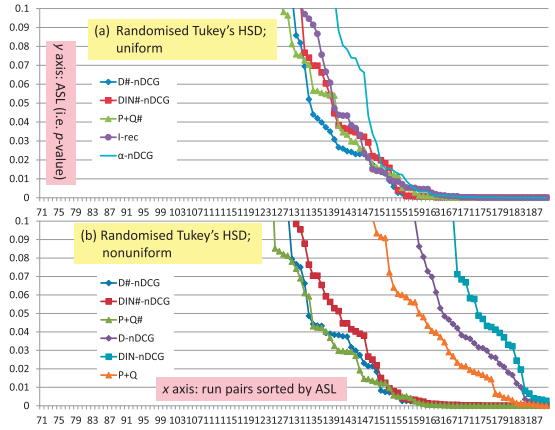

Fig. 9 ASL curves based on the randomised Tukey's HSD.

5.2 Evaluation Toolkit

For computing all evaluation metrics, we used the *NTCIREVAL* toolkit [20][*10]. The only exception was α-nDCG: we 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

Figure 8 and **Fig. 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.

Table 2 and **Table 3** cut Fig. 8 and Fig. 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.

[*10] Available at http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html.

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			

First, by comparing the bootstrap and the randomised Tukey's HSD results (i.e., Fig. 8 vs. Fig. 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 (Fig. 9 (b)) but not with the bootstrap test (Fig. 8 (b)).
- The randomised Tukey's HSD is substantially more conservative than the bootstrap test, as it is clear from the contrast between Fig. 8 and Fig. 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 (Table 2 (a) and Table 3 (a)).

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

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

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 Concordance test results. For each metric pair, the higher score is shown in bold. Disagreements are shown in parentheses. Significant differences according to the two-sided sign test are indicated by †'s ($\alpha = 0.05$) and ‡'s ($\alpha = 0.01$).

(a) Gold standard: 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) Gold standard: Ef-P ("effective 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) Gold-standard: I-rec AND Ef-P			
	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)

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 Fig. 8 and Fig. 9 and Table 2 and Table 3, it can be observed that:

- DIN♯-nDCG and P+Q♯ are comparable to I-rec and D♯-nDCG in terms of discriminative power (Fig. 8 and Fig. 9)[*11].
- α-nDCG may be slightly less discriminative than the above best metrics (Fig. 8 (a) and Fig. 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 Fig. 8 (b) and Fig. 9 (b));
- 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 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 [30] 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 sensitive to swaps near the top ranks. 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 rationales behind them (see Section 3).

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

5.4 Concordance Test Results

Table 6 show the concordance test results for α-nDCG, D♯-nDCG, DIN♯-nDCG and P+Q♯ computed using the algorithm shown in Fig. 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 concordance 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 concordance score for α-nDCG is only .597, that for D♯-nDCG is .995. This means that,

[*11] DIN♯-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.

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. This difference is statistically significant according to a two-sided sign test at $\alpha = 0.01$. The relative results can be summarised as follows:

(1) In terms of diversity (Part (a)), D♯-nDCG, DIN♯-nDCG and P+Q♯ significantly outperform α-nDCG; P+Q♯ significantly 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)).

(2) In terms of effective relevance (Part (b)), D♯-nDCG and DIN♯-nDCG significantly outperform α-nDCG, and also significantly outperform P+Q♯; and therefore D♯-nDCG and DIN♯-nDCG are the winners.

(3) In terms of both diversity and effective relevance (Part (c)), D♯-nDCG, DIN♯-nDCG and P+Q♯ significantly outperform α-nDCG; DIN♯-nDCG significantly outperforms P+Q♯ in addition; 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)). In particular, P+Q♯ is highly consistent with I-rec in Part (a) because the raw P+Q values are relatively low: recall that a single ranked list is highly unlikely to achieve a very high performance value for all intents at the same time (Section 3.2). When the raw P+Q value are low, the impact of I-rec on the P+Q♯ is high, due to a linear combination that is similar to 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. Nevertheless, the concordance test results are valuable because they allow quantitative comparisons and show exactly how often one metric outperforms another with real data. According to our results, DIN♯-nDCG is the best metric that takes both diversity and effective relevance into account (from Part (c)). Moreover, note that both D♯-nDCG and DIN♯-nDCG significantly outperform α-nDCG from the viewpoint of rewarding diversity (Part (a)) *and* from the viewpoint of rewarding effective relevance (Part (b)).

Note also that our concordance test is applicable to *any* pair of evaluation metrics provided that an appropriate gold-standard metric that represents a desirable property 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 used at TREC handles neither.) We also proposed the concordance test for comparing the intuitiveness of a given pair of metrics quantitatively. Our main experimental findings are:

(a) In terms of *discriminative power* [17], [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;

(b) In terms of the *concordance test* which quantifies the agreement of a diversity metric with a gold standard metric that represents a basic desirable property, DIN♯-nDCG is superior to other diversity metrics in its ability to reward both diversity and relevance at the same time. Moreover, both D♯-nDCG and DIN♯-nDCG significantly outperform α-nDCG in their ability to reward diversity, to reward relevance, and to reward both at the same time.

In addition, we demonstrated 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. We therefore recommend the former approach for significance testing when a set of runs is available for evaluation.

Finally, limitations of the present study include the following:

(1) As was discussed in Section 3, DIN♯-measures and P+Q♯ 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 undertand and to compute. Recall that computing α-nDCG requires a greedy approximation of the ideal list.

(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. [26] 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 [23]. (But recall that our experiments involve 68 informational intents and 31 navigational intents as shown in Table 1.) We will conduct similar experiments using the NTCIR INTENT test collections which come with intent probabilities obtained through assessor voting [24], [29].

(4) While the proposed metrics leverage the explicit knowledge of whether each intent is informational or nagivational, 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.

References

[1] Agrawal, R., Sreenivas, G., Halverson, A. and Leong, S.: Diversifying Search Results, *Proc. ACM WSDM 2009*, pp.5–14 (2009).

[2] Brandt, C., Joachims, T., Yue, Y. and Bank, J.: Dynamic ranked retrieval, *Proc. ACM WSDM 2011* (2011).

[3] Broder, A.: A Taxonomy of Web Search, *SIGIR Forum*, Vol.36, No.2, pp.3–10 (2002).

[4] Carterette, B.: Multiple testing in statistical analysis of systems-based information retrieval experiments, *ACM TOIS*, Vol.30, No.1, Article No.4 (2012).

[5] Chapelle, O., Metzler, D., Zhang, Y. and Grinspan, P.: Expected Reciprocal Rank for Graded Relevance, *Proc. ACM CIKM 2009*, pp.621–630 (2009).

[6] Clarke, C.L., Craswell, N. and Soboroff, I.: Overview of the TREC 2009 Web Track, *Proc. TREC 2009* (2009).

[7] Clarke, C.L., Craswell, N., Soboroff, I. and Ashkan, A.: A Comparative Analysis of Cascade Measures for Novelty and Diversity, *Proc. ACM WSDM 2011* (2011).

[8] Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A.,

Büttcher, S. and MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation, *Proc. ACM SIGIR 2008*, pp.659–666 (2009).
[9] Clarke, C.L., Kolla, M. and Vechtomova, O.: An Effectiveness Measure for Ambiguous and Underspecified Queries, *Proc. ICTIR 2009*, pp.188–199 (2009).
[10] Dou, Z., Hu, S., Chen, K., Song, R. and Wen, J.-R.: Multi-dimentional Search Result Diversification, *Proc. ACM WSDM 2011*, pp.475–484 (2011).
[11] Jansen, B.J., Booth, D.L. and Spink, A.: Determining the Informational, Navigational, and Transactional Intent of Web Queries, *Information Processing and Management*, Vol.44, pp.1251–1266 (2008).
[12] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol.20, No.4, pp.422–446 (2002).
[13] Leelanupab, T., Zuccon, G. and Jose, J.M.: A Query-basis Approach to Parametrizing Novelty-biased Cumulative Gain, *Proc. ICTIR 2011*, pp.327–331 (2011).
[14] Macdonald, C., Wang, J. and Clarke, C. (Eds.): *Proc. the Diversity in Document Retrieval 2011 Workshop*, University of Glasgow (2011).
[15] Rafiei, D., Bharat, K. and Shukla, A.: Diversifying Web Search Results, *Proc. ACM WWW 2010*, pp.781–790 (2010).
[16] Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, *Proc. AIRS 2006* (*LNCS 4182*), pp.374–389 (2006).
[17] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *Proc. ACM SIGIR 2006*, pp.525–532 (2006).
[18] Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, *IPSJ Trans. Databases*, Vol.48, No.SIG9 (TOD35), pp.11–28 (2007).
[19] Sakai, T.: On the Properties of Evaluation Metrics for Finding One Highly Relevant Document, *IPSJ Trans. Databases*, Vol.48, No.SIG9 (TOD35), pp.29–46 (2007).
[20] Sakai, T.: NTCIREVAL: A Generic Toolkit for Information Access Evaluation, *Proc. FIT 2011* (*Volume 2*), pp.23–30 (2011).
[21] Sakai, T.: Evaluation with Informational and Navigational Intents, *Proc. WWW 2012*, pp.499–508 (2012).
[22] Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, *Proc. EVIA 2008*, pp.30–41 (2008).
[23] Sakai, T. and Song, R.: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, *Proc. ACM SIGIR 2011*, pp.1043–1052 (2011).
[24] Sakai, T. and Song, R.: Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT Task, *Information Retrieval*, to appear (2013).
[25] Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems, *Foundations and Trends in Information Retrieval*, Vol.4, pp.247–375 (2010).
[26] Sanderson, M., Paramita, M.L., Clough, P. and Kanoulas, E.: Do user preferences and evaluation measures line up?, *Proc. ACM SIGIR 2010*, pp.555–562 (2010).
[27] Santos, R., Macdonald, C. and Ounis, I.: Intent-Aware Search Result Diversification, *Proc. ACM SIGIR 2011*, pp.595–604 (2011).
[28] Smucker, M.D., Allan, J. and Carterette, B.: A Comparison of Statistical Significance Test for Information Retrieval Evaluation, *Proc. ACM CIKM 2007*, pp.623–632 (2007).
[29] Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q. and Orii, N.: Overview of the NTCIR-9 INTENT Task, *Proc. NTCIR-9*, pp.82–105 (2011).
[30] Yilmaz, E., Aslam, J. and Robertson, S.: A New Rank Correlation Coefficient for Information Retrieval, *Proc. ACM SIGIR 2008*, pp.587–594 (2008).
[31] Zhai, C., Cohen, W.W. and Lafferty, J.: Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval, *Proc. ACM SIGIR 2003*, pp.10–17 (2003).

Tetsuya Sakai was born in 1968. He received a Master's degree from Waseda University in 1993 and joined the Toshiba Corporate R&D Center in the same year. He received a Ph.D. from Waseda University in 2000 for his work on information retrieval and filtering systems. From 2000 to 2001, he was a visiting researcher at the University of Cambridge Computer Laboratory. In 2007, he became Director of the Natural Language Processing Laboratory at NewsWatch, Inc. In 2009, he joined Microsoft Research Asia, where he has worked ever since. He has received several awards in Japan, mostly from IPSJ. He chaired IPSJ SIG-IFAT from 2009 to 2010, and is currently an editor-in-chief of IPSJ TOD.